DOCUMENT RESUME

ED 166 725 CS 204 715

AUTHOR	Levine, Arvin
TITLE	English Intonation and Computerized Speech Synthesis. Technical Report No. 287.
PUB DATE	Jul 77
NOTE	308p.; Ph.D. Dissertation, Stanford University
EDRS PRICE	MF-$0.83 HC-$16.73 Plus Postage.
DESCRIPTORS	*Acoustic Phonetics; *Artificial Speech; *Computational Linguistics; Discourse Analysis; Higher Education; Language Research; Oral English; Phonetics; *Phonology; *Semantics; Sentence Structure; Speech; *Syntax
IDENTIFIERS	*Oral Language

ABSTRACT

This work treats some of the important issues encountered in an attempt to synthesize natural sounding English speech from arbitrary written text. Details of the systems that interact in producing speech are described. The principal systems dealt with are phonology (intonation), phonetics, syntax, semantics, and text-view (discourse). Technical details of a full-speech synthesis system are included; appendixes provide loudness and fundamental frequency graphs for selected sentences, data on peak pitches from four separate readers who read the same paragraph, and the audio text of a lesson from an elementary college course in logic. (Author/GT)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED166725

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ENGLISH INTONATION AND COMPUTERIZED SPEECH SYNTHESIS

by

Arvin Levine

TECHNICAL REPORT NO. 287

July 29, 1977

PSYCHOLOGY AND EDUCATION SERIES

Reproduction in Whole or in Part Is Permitted for
Any Purpose of the United States Government

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

CS204 715

ENGLISH INTONATION AND COMPUTERIZED SPEECH SYNTHESIS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF LINGUISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Arvin Levine
July 1977

© Copyright 1977

by

Arvin Levine

4

ACKNOWLEDGEMENTS

There are people and there are organizations that I would like to thank. The organizations, being less delicate, are easier to give recognition to, and they are: Stanford University, in particular the Department of Linguistics and the Institute for Mathematical Studies in the Social Sciences; the National Science Foundation; and the Department of Health, Education and Welfare, Bureau for Education of the Handicapped (BEH). The University in general and the linguistics department have supported my graduate existence. The Institute provided a project to work on and the various government agencies have contributed financial support for the work.

There are many different people who, to greater and lesser extents, have been responsible to greater and lesser extents for my completing this dissertation. I would like to for my completing this dissertation. I would like to thank them at this time and be done with it.

There are some people who deserve the infamy of being singled out by name for having helped me along. My advisors and reading committee are prime candidates for recognition of this sort. First and foremost I want to thank Will Leben for involving me in this research and for his generous help advice and guidance. I probably can't thank him enough, so I will stop trying. Terry Winograd and Tom Wasow provided important comments that straightened out a lot of what I wanted to say. Whatever remains gibberish is my responsibility alone. I thank Pat Suppes for looking over my shoulder and moderating some of the unbridled optimism I have been wont to exhibit.

Among my colleagues at the Institute, I want particularly to mention Bill Sanders, Bob Laddaga, Mike Hinckley and Bob Smith for the encouragement and moral support that I got from them. There are a lot of other nice people at the Institute, too numerous to mention individually. Within my department, Dick Oehrle, Joel Katz, Debby Ohsiek, Jim Gee, Clara Bush and Dorothy Huntington have helped me out, in one way or another.

Somehow I found time to associate with people not directly involved in the sort of work I was doing. For them, a quick "thanx" for whatever: Ann Fogle, Lou Steinberg, Nachum, Shulamit and Erga Dershowitz, Yosef Tuvia haCohen, David Schlossman, Ray Finkel, Mishpachat Aumann, Jack Hausner, John and Judy Rechenberg; a dozen among a cast of thousands.

Finally, I want to mention my parents, Saul and Pauline, and siblings, Gavrielle and Myron. They have contributed to my work in ways that I will probably never understand. I must also thank and bless the Holy One, Blesed be He, under whose protection I have reached this momentous occasion.

ENGLISH INTONATION AND COMPUTERIZED SPEECH SYNTHESIS

Arvin Levine
Stanford University, 1977

ABSTRACT

This work treats some of the important issues encountered in an attempt to synthesize natural sounding English speech from arbitrary written text. Details of the systems which interact in producing speech are described. The principal systems dealt with are phonology (intonation), phonetics, syntax, semantics, and text-view (discourse). While further research will advance the understanding necessary for high quality speech synthesis, it will be demonstrated that the work described below forms a reasonable basis for such future research.

Approved for publication:

By ____________________________
For Department of Linguistics

By ____________________________
Dean of Graduate Studies

Table of Contents

Chapter Page

Section

List of Figures x

I. Introduction 1

I.1 The problem 2

I.2 A note on terminology 25

I.3 Some limitations of the research 27

I.4 Overview of the dissertation 30

I.5 The computer environment of this research 40

II. English as a tone language 46

II.1 Preliminary discussion 47

II.2 The autosegmental hypothesis 57

II.3 The correspondence of tone languages and intonation 61

II.4 Tone groups 69

II.5 Emphatic stress: a result of tonal elaboration . . 79

II.6 Another derivative result 81

III. Experimental results 92

III.1 Simple sentences 92

III.2 More complicated sentences 96

III.3 Some complicated simple sentences 106

III.4 Some simple questions 119

III.5 Phonology and the Real World of Phonetics . . . 124

III.6 Other ideas that have been examined 141

IV. Metrics and duration 149

IV.1 Some elementary facts 151

IV.2 Possible duration assignments 156

IV.3 Data 165

IV.4 Our current approximation to duration assignment 178

IV.5 Deficiencies and alternatives 179

V. The syntactic component of the model 185

V.1 Some complicated parsers 187

V.2 The current parser 196

VI. Semantics and text-view 212

VI.1 Semantics 213

VI.2 Text-view 227

VII. Technical details of a full speech synthesis system . . 237

VII.1 TEXTSP, a prototype text-to-speech system . . . 237

VII.2 The punctuation scanner 243

VII.3 Specifying the syntax 245

VII.4 Prosodic specification generation 251

VII.5 The command translator 256

VII.6 Prosodic contour imposition 258

Appendix I

Fundamental frequency and loudness graphs 259

Appendix II

The North wind and the sun 266

Appendix III

Text of the lesson with comments 273

References 289

List of Figures

Figure Page

1. Preferred and confusing terms for concepts related to speech synthesis. 27

2. Some two and three word proper names and sentences. 32

3. An abstract example. 33

4. Block diagram of the computer system. 41

5. Schematic representation of the audio lexicon. 46

6. Description of a pitch contour. 51

7. A hierarchy of concrete tones. 51

8. An elaboration hierarchy for a tone group. 55

9. Description of a pitch contour. 56

10. Description of a pitch contour. 57

11. Bernstein's schema relating language and music. 73

12. A Mozart symphony as elaboration of one phrase. 75

13. Examples of non-reduction before a deletion site. . . . 84

14. Selkirk's monosyllable rule. 85

15. A low stress NON-mid tone retains stress. 86

16. A low stress mid tone reduces stress. 87

17. A more complicated example of stress reduction. 87

18. Two word sentences and proper names. 93

19. Three and four word sentences and proper names. 94

20. Complex sentences. 97

21. Complicated simple sentences. 107

22. Questions. 120

23. Segmental durations for selected words. 153

24. Sample durations arranged by word class. 166

25. Sample durations arranged by word class. 170

26. Percentage durations arranged by word class. 171

27. Percentage durations arranged by word class. 172

28. Percentage durations arranged by word class. 173

29. Durations of words for selected sentences. 176

30. Duration assignments. 178

31. Duration relationships abstracted from Gaitenby(1965) . . 180

32. Some samples of rearrangement of lexical stress patterns. . 183

33. Overview of our surface parser. 188

34. Part 1: Summary of constructions parsed. 200

35. Part 2: Summary of constructions parsed. 201

36. Part 3: Summary of constructions parsed. 203

37. Some observed pitch contours. 215

38. Some schematically shown effects of focus shift. 220

39. An example of shifted focus. 221

40. Some sentences with syntactic bracketing and amplified for semantic information. 227

41. Constituent structure of several paragraphs 229

42. Text-view structure: percentage difference of observed length of utterance from length prediction using only syntax and lexicon. 233

43. The components of the TEXTSP system. 238

44. Flowchart/outline of the TEXTSP system. 240

45. Component diagram for a revised text-to-speech system. . . 242

46. CF Specifications for the prosodic control language. . . 246

47. Explanation of terminal symbols. 247

48. Sample bracketed strings. 249

49. Prosodic assignments corresponding to the phrase structure rules. 252

50. Tree structures showing the assignments of prosodic parameters to bracketed strings. 255

51. Summary of prosodic assignments in sample cases. 256

Chapter I

Introduction

This work represents an attempt to deal with some of the important issues encountered in an attempt to synthesize natural sounding English speech from arbitrary written text. A naive impression of sentential intonation might lead one to say that stress is the most important feature of spoken English. Yet it has generally been recognized that the phenomenon known as stress in English has no direct instrumental correlate[1]. Rather it seems to be related to a combination of pitch relationships, durational relationships and amplitude relationships. We must therefore consider stress to be simply one reflex of the total intonational system. The bulk of this work reflects the de-emphasis of stress per se with a corresponding emphasis on systemic and prosodic relationships.

The work reported in this dissertation has been carried forward with two objectives. One objective has been to select a basis and carry out experiments concerning theoretical aspects of intonation in English (analysis and synthesis). The other object has been an attempt to construct a speech synthesis system (using word concatenation with intonation modification) that can complement a computer aided instruction system or which has applications as a "text-to-speech" system for the blind.

---

[1] Many researchers have pointed this fact out, notably Fry (1958), Bolinger (1965, 1972), Ohala (1970).

I.1 The problem

The problem of synthesizing natural sounding English speech has been approached from many directions. This research bears marks of two approaches in particular, the linguist's approach and the computational approach. These approaches are best exemplified by the different formulations of the problem given below. We return to a discussion of these dichotomies afterwards.

I.1.1 A linguist's approach

The following two descriptions of the problem relate principally to interests commonly associated with theoretical linguistics.

I.1.1.1 Statement one

Describing the pitch contours, much less the more general intonational structure, for sentences or phrases of English seems at first glance to be a hopeless task. The immediate problem facing the linguist is that of describing the continuum of data in some discrete fashion. Another aspect of the same problem that has been under discussion by linguists for about fifty years is the definition of the phoneme. The basic theoretical justification for the use of phonemes is explicated by Swadesh (1934:1968)[2], and, practical reasons for defining the phoneme are stated by Twaddell (1935:1968).

---

[2] Many of the important early papers on the subject of the phoneme are reprinted in readers by either Joos (1957:1968) or Makkai (1972). Besides Swadesh and Twaddell, who are cited here, there are many other readings bearing on the history of the phoneme to be found in those readers.

> The phonemic principle is that there are in each language only a limited number of elemental types of speech sounds, called phonemes, peculiar to that language; that all sounds produced in the employment of the given language are referrable to its set of phonemes; that only its own phonemes are at all significant in the given language. (Swadesh: p.32)

> If a satisfactory definition of the phoneme as an abstractional fiction can be worked out, that definition would have two advantages: (1) we could use the phoneme as a terminological convenience ...; (2) we should not have to use .... the promissory notes of the laboratory .... (Twaddell: p.67)

We need to be able to make an abstraction that will perform the same function for intonation that the phoneme does for segmental phonology. The possibility that phonemes are not basic units but are defined in terms of distinctive features does not complicate the analogy we would like to establish between phonemes and possible tonal units. Suppose that the units of tonal analysis which we select could be further defined in terms of some other tonal unit. The basic issue involved would be the usefulness of the different sets of units for capturing generalizations about language. It might even be reasonable to allow both types of unit to coexist.

Several different basic units are conceivable. One possible solution is to postulate a basic unit which is itself continuous. For example, we might consider that the set of basic units available to a human language consists of levels of pitch, e.g. low, mid, high, and, pitch changes, e.g. fall, rise, fall-rise, rise-fall, fall-rise-fall, etc. These could then be combined either by virtue of universal principles or in language particular ways to generate the observed pitch contours for the language.

One approach using levels and changing pitches as basic units applied to French intonation can be found in Martin (1974, 1975). His approach, as well as the one defended here, uses the syntactic structure of the utterance to guide the assignment of intonational units to the structure. In both theories the intonational contour is defined in terms of the units employed. An important difference between the two theories is in the nature of the units which are used to describe the contours. His system uses a set of units which can be described as level pitches ending in either a rise or fall, the rise or fall being strong or weak and either 'extreme' or 'median'. Martin uses a set of three binary features ([+ rising], [+ strong], [+ extreme]) to define the eight possible minimal "changing pitch" units of his system. Our statement that the basic unit for Martin is a pitch change is nonetheless true since only the "changing pitch" units, and not the features, play an interesting role in the system he develops for intonation assignment. The net effect of the distinctive features is the representation of "changing pitches" of various sorts in formally accessible terms. This work is a serious attempt to describe intonation in linguistic terms which uses as basic elements pitch change descriptors.

Martin's approach falls within the general paradigm of intonation studies that might be called "the English School" as typified by Crystal (1969), Halliday (1973), O'Connor and Arnold (1961), Palmer (1922) and others[3]. This school analyzes pitch contours into a nucleus preceded (optionally) by a prehead and head and followed by a tail. Each of these units consists of a changing-tone, which may be a level tone. These changing tones correspond to the distinctions which Martin made enlarged with a repertoire of more complex changes, e.g. rise-fall-rise. The basis for admitting a changing-tone to the basic repertoire is the possibility of linguistic contrast with another tone in the same position.

An independent approach which applies similar units to English intonation is that of the M.I.T. group led by J. Allen. For example, we have a statement on 'choice of parameters' from O'Shaughnessy (1976, pp. 71-2):

> In light of the findings about which F0 movements were most readily distinguished perceptually and which F0 gestures were produced ..., this thesis accepted the motions of F0 (rising, falling, or level) on each syllable as the most relevant aspect of F0 patterns.

On this account, pitch accents are defined in terms of the degree of rise or fall of specific parts of the contour. These accents are determined by the different syntactic and semantic structures (both neutral and 'stress' patterns) of an utterance. Both the Martin and O'Shaughnessy accounts concentrate on pitch changes in English Intonation.

---

[3] A good review of the literature is in Crystal (1969).

Many approaches to tone language phenomena (as opposed to intonation studies) have utilized changing pitches as basic units. Common orthography also usually selects this kind of unit to describe the tone phenomena as written, i.e., the various accents, acute, grave and circumflexed, can indicate either different levels of tone or tone contours.

Alternatively, we might insist on a unit which implies a discrete representation of the observed pitches, allowing only pitch levels. In some languages, Japanese for example[4], the pitch level account needs little emendation in order to be compared to the phonetic data since the use of pitch is limited to relationships between reasonably level pitches, even on a phonetic account. For other languages we will need to complicate any phonological account which has as units only level pitches to account for the non-level aspects of observed pitch contours. Even in these languages (with largely non-level pitch contours) there will probably be some intonational phenomena in which the abstract phonological account corresponds fairly closely to the phonetic data. Singing, and the 'vocative chant' intonation, described in Liberman (1975) and Leben (1976) provide examples of the simple realization of an abstract account (using level tones) as a phonetic account. Tone levels have been used to describe English intonation by several researchers, notably Goldsmith, Leben, and M. Liberman. Further discussion of their proposals can be found below, Section II.2. While the tone levels are the significant data for this theory, we also generate the continuous pitch changes of an utterance's contour by a limited set of "low level" phonetic rules. These rules predict the nature of the pitch contours between the distinct pitch levels which are the generated by the phonological component.

---

[4] Cf. Ohsiek (1976).

This research effort has adopted the latter approach, describing intonation in terms of level tones. We have selected a set of basic levels and developed phonological rules that manipulate these levels to achieve an abstract account that corresponds to the observed data. As we will attempt to show later, this approach yields a (linguistically) more fruitful theory of intonation. In addition, we have discovered phonetic rules that mediate between the output of the phonology that we use and the actual phonetic outputs required for speech synthesis.

The basis for the prosodic manipulations that constitute the core of this dissertation is the "tone group" hypothesis. Essentially the tone group hypothesis states that all pitch contours can be analyzed as having an underlying melody which is composed of distinct tones. This melody is elaborated and then mapped onto (or associated with) the segmental string. The underlying melody is a theoretical construct in the sense that it almost never appears on the surface as such. There are several reasons for utilizing this level of abstraction, and these will be discussed later. Just as the underlying melody is a theoretical entity, the underlying tones which compose the melody exist only as theoretical constructs. That is, our discussion will involve only three underlying tones, low (L), mid (M), and high (H). These underlying tones will appear as the surface (phonetic) pitches associated with a sentence or phrase only in the simplest of cases. In general the elaboration of the melody will produce hierarchical structures composed of these tones which will have definite phonetic realizations as specific pitches. In addition we will propose phonetic mechanisms to reproduce the entire pitch contour as a function of these pitches.

There are two sources of motivation for this tone group hypothesis. One comes from music and the other from studies of actual tone languages. The similarity of speech and song suggests that some form of musical analysis can bear on speech analysis. The concepts of underlying melody and tones outlined above, and developed further below, correspond to the use of melody and notes in musical analysis. Consideration of possible universals of human speech suggest that tone languages should not be intrinsically different in their repertoire of speech elements from non-tone languages[5]. The basic concept drawn from comparison with tone languages, the autosegmental hypothesis, will be discussed in greater depth later. It concerns the connection between the tonal (pitch) assignments and the phonetic segments of a sentence or phrase.

---

[5] In this respect, an interesting comment was made by Bolinger, in a presentation at Stanford University (1977), that universals of intonation are, in a sense similar to universals of facial expression, etc. While he was referring strictly to intonation, we can generalize the notion to say that the same tonal mechanisms (e.g. duration, pitch and loudness) are universally available, whether a particular language uses them intonationally or segmentally.

The difference between our approach and those using units composed of changing pitches lies in the generalizations which each set of units can allow us to capture. While both may turn out to be empirically verified since there is no reason, in principle, why each can not describe the other, we feel that our approach will yield more interesting (to linguists who are interested) statements of the general principles underlying English intonation.

I.1.1.2 Statement two

A problem for the description of English intonation is the difficulty of combining in a single, unified description all of the sources of intonational effects. Even looking at only one of the many intonation parameters is complicated. Klatt (1976) has identified eleven factors that relate to the durational structure of a sentence. These factors can be collected under three type headings:[6]

1) Word (and sub-word) level phonetic effects. These effects include consonant incompressibility, segmental interactions, inherent phonological duration and word final lengthening.

2) Syntactic and semantic phrase level effects. These include phrase structure lengthening, emphasis and semantic novelty.

3) Discourse and extralinguistic effects. These include the psychological and physical state, the

---

[6] Klatt, in his article, has seven type headings, but his purposes are different from mine.

speaking rate and, among discourse phenomena, position in a paragraph.

More generally, we can enumerate all the factors affecting intonation under these same three type headings. For example, it is well known (Cf. Hombert 1973a,b,c, Hyman 1973, O'hala 1973 and other references cited there) that different vowels and consonant/vowel aggregates are responsible for identifiable contributions to the pitch contour. Similarly, recognizable contributions from the other two types of interactions can be identified.

The research presented here assumes that the first type of effects, that of word and sub-word level phenomena are already known. In the actual speech synthesis system this assumption is incorporated by performing the synthesis operations on pre-stored words which are already specified for pitch, duration and loudness. If speech synthesis from phoneme concatenation were desired, instead of synthesis from word concatenation, then an additional module incorporating research results from this area of word and sub-word phonology and phonetics could be incorporated[7]. We obtain our words by actual recording of specific words in the majority of cases, on the hypothesis that a lexical recording of a word inherently contains most of the appropriate phonetic information, particularly for word internal structure. We have not performed research on the nature of word boundary effects on the phonetic segmental (as opposed to intonational) structure of utterances. While it is certainly true that the phonetic information for these effects is not contained in the lexical recordings we are using, many of the effects can be ignored in practice under the aegis of generating "distinct speech". The theoretical status of a lexical intonation to the 'prosodied' intonation will be discussed further below, Section II.1.

---

[7] More discussion of the importance of word level phenomena to our synthesis system can be found with the details of implementation, Chapter VII.

We deal with the third type of effects, those relating to extralinguistic and discourse phenomena, by assuming constancy of effect for our corpus and postponing close examination of them for future research efforts. Thus, while speaking rate may vary with a speaker's psychological state, we can ignore that effect, for the time being, by assuming that his emotional state is relatively constant, both when analyzing recorded data and for purposes of speech synthesis. A large part of our study is concerned with the second type of intonation effect, those of a syntactic and semantic nature.

I.1.1.3 Synthesis one

The two approaches outlined above indicate the division of this research into one portion of strictly phonological/phonetic interest and one portion of syntactic/semantic interest. The central concept in the first statement of the problem was addressed solely towards phonological considerations. There the question of interest involved selecting a basic phonological unit. This outlook on the problem allows us to transcend the immediate question of English intonation and to take a broader perspective regarding pitch in language in general, both in languages where the tonal phonology plays a more central role, and those in which pitch only enters the phonology at a more surface level. This viewpoint gives us the basis to deal with the second version of the problem of intonation. We can describe the relationship of syntax and semantics to phonology by explicitly linking the phonological constructs to syntactic and semantic units. The fundamental link for us is the use of surface syntactic structures to guide the tonal elaboration. The link to semantic content, in this theory, is mediated by the syntactic structure. We can also motivate results which are not obviously cases of intonation from our general intonation generation principles. An example is our explanation of vowel reductions and contraction of minor lexical items[8]. Previous attempts to account for the data directly involved some degree of obtuseness whereas our account, given our overall approach, is much simpler and more straight-forward.

I.1.2 A computational approach

The following problem statements reflect the interests of researchers normally associated with computational linguistics, especially speech synthesis.

I.1.2.1 Statement three

An important question in artificial intelligence research is the range of tasks for a given model of language use. In opposition to the possibility that there exists a unique representational schema which can be efficiently used for any computational need, it seems reasonable to believe that the usefulness of a given representation will be dependent on the problem which is being addressed. A representation which is especially suited to doing arithmetic may be unsuitable (and inefficient) for performing symbolic calculations. A program whose function is to answer questions based on a (potentially) complex sequence of logical deductions from some initial hypotheses will be most efficient with a representation system that closely resembles the logical calculus. A different representation might be most reasonable for a program which attacks a different problem, for example a semantic network would be a reasonable representation for a program that tries to do "common sense" reasoning, such as retrieving (previously stored) relevant information on a common theme.

---

[8] see Section II.6.

The different representations (and programs using them) that are currently being developed for different artificial intelligence projects reflect different views about some aspects of the mental makeup of human beings. In that light, the choice between a multiplicity of structures, each dedicated to a relatively specific purpose, and a unique structure that does everything corresponds to a choice between a mental model specifying skill areas as distinct from each other (and possibly innate) and a model in which there is a single, multipurpose mental facility (which could not have innate capabilities for specific functions[9]). This argument is not limited to the AI community. For example, Chomsky (1975 p.20, and elsewhere) can be read as arguing that there is a distinction to be made between the 'language faculty', which he claims to be species specific, and a more general problem-solving ability which may well exist in other, non-human species, as well. Thus he seems to be arguing for a diversity of mechanisms and against the idea of a unique representational scheme.

---

[9] Innateness implies specificity of function. Of course, even on this second model, there could be innate mental features which allow the human brain to develop its general facility to specifically human skills.

From this perspective, our problem is to identify the aspects of language use which need to be incorporated in a model of the intonational system of English. In the Sound Pattern of English (SPE) the different linguistic components are identified as interacting in a completely modular way, where the output from one component becomes the input to the next. The linguistic interactions responsible for intonation can provide evidence about the nature of a representation scheme in the same way that the hypotheses about the interactions between syntactic and segmental phonological rules in SPE provided evidence about the nature of a representation for language use.

If intonation were a purely phonetic matter, this problem would be trivial. We would only need to model the phonetic data relevant to intonation (perhaps on a statistical basis) while disregarding any other linguistic systems (syntax, etc.). Such an intonational model would resemble a model of, e.g. phoneme concatenation. The problem is not trivial since there are interactions between most of the language sub-systems and intonation. Many people have noticed that the semantics and pragmatics of an utterance influence (alternatively: are influenced by) the intonation contour of the utterance. One example (reported in Bolinger 1972) which illustrates this interaction is as follows. Two different meanings can be imputed to the sentence, "Do you want coffee, tea, or milk?" depending on whether a (so-called) comma or "disjunctive" intonation separates the words, 'coffee' and 'tea' or not. This intonation is a steep rise in pitch represented below as 'LH' (Low tone to High tone). The last word of this contour has a fall, represented 'HL' (High tone to Low tone) typical of wh- questions (see below). The non-comma or neutral intonation is a low tone until the final word, 'milk', which is intoned with a 'LH' contour due to the normal rise at the end of a yes/no question.

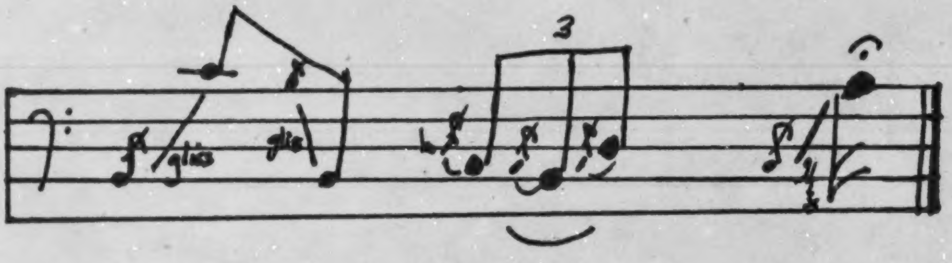
	Do you want coffee,	tea or	milk?	
disjunctive intonation:	L H	LH	HL	(chose one)
normal intonation:	L	L	LH	(anything at all)

If the comma intonation is present then the sentence is offering a choice of precisely one of the three beverages available. An appropriate answer to the comma intonation would be, "I'll have milk, thank you." There is a supposition that you want something to drink. If the non-comma intonation is present then the sentence is asking if you want something (anything) to drink. An appropriate answer to this intonation would be, "No, thank you."

Another example of the interaction of intonation and semantics/pragmatics is due to Liberman and Sag (1974). They give the following "conversation" (p. 422, their example 10) illustrating what they call the contradiction contour.

Mark: Hey Ivan, how about on your way to school this morning you drop off my pet whale at the aquarium?

Ivan: (Kazoo or slide whistle; ...)



They contrast a range of possible texts that could fit into the semantic/pragmatic context created by using this intonation contour (their examples 11 and 12) which we partially reproduce below.

Appropriate texts:

|You don't have a pet whale.!
|I'm not taking orders from you anymore!

Inappropriate texts:

|I'm more than happy to take him along!
|I'm not very fond of that animal!

The point is clear that the contradiction contour is the crucial linguistic fact determining the appropriateness of the various texts since, with other contours, it would be possible to consider all of the texts equally appropriate. There is a definite relationship in this case between the intonational component and the semantic/pragmatic components of a grammar of English.

Another possible answer to the general problem of the interaction of intonation with other linguistic systems would claim that only some systems interact with intonation. For example, it is conceivable to claim that the semantic structure of an utterance is crucial to the intonational contour but the syntactic contribution is minimal or even non-existent. In particular, a claim could be made that the semantics determines the stress and thus some parts of the overall intonation, but otherwise the intonation is essentially random or at least underdetermined. This view is close to classical positions (e.g. Crystal (1969)) in which only the 'head' of a phrase and other distinguished or prominent elements in the phrase get actual tone assignments. Goldsmith's (1974,5) autosegmental system (and Leben (1976) following the same formalism) also fits into this model of semantic/phonological but non-syntactic prominence. In this type of system, a prominent element is assigned a prominent tone. The other tones are assigned by phonological rules. In Leben, the main tone rules only assign tones to stress bearing syllables, other syllables receive their tones from phonetically derived (or low level phonological) tone spreading rules. While the prominence could be syntactically motivated (say by some variant of the Nuclear Stress rule, see Bresnan (1971)) the assumption of these researchers has been that prominence is either a semantic or phonological property of phrase and words, respectively.

Our position is that both the semantic and syntactic structures of an utterance play important roles in determining its intonation. We have cited, above, some of the evidence linking the semantic and intonational structures; our experimental evidence (see below Chapter III) strongly indicates that the syntax is also linked to the intonational structure. We feel that it is worthwhile to isolate the individual components so that the contributions of each can be identified. The methodology we employed was to start with cases in which the semantic contribution could be assumed not to interfere with the other systems. Then, having identified the basic syntactic results, we could address the semantically related cases and observe those interactions, factoring for the syntax. By limiting the interactions of the syntax and semantics in this research, we have been able to make some fairly strong claims about the amount and sorts of information needed to synthesize speech by word concatenation and modification.

I.1.2.2 Statement four

Another computational approach to speech synthesis can be characterized as an engineering approach. According to this approach an important discovery would be a statistically verified (or verifiable) formula or algorithm for doing the synthesis processing. It would be an advantage for this approach if the algorithm were as economical as possible, but it is also conceivable that the most economical algorithm from this point of view might not correspond with the linguistic (or philosophical) notion of economy.

There is embodied in this view of the problem the suggestion of a "brute force" method of solution. Some researchers have relied on statistical methods to determine relationships between different components of speech. This approach is typified by the following, taken from Ohman and Lindqvist[10].

> Our experience so far with the analysis of pitch contours suggests that the essentials of these speech patterns may be summarized in terms of of simple step function complexes ....

While this is not necessarily in conflict with the desire for linguistic relevance supported here, such an attitude easily leads to focusing on computationaly "reasonable" approaches rather than linguistically interesting ones. For example, Rabiner and Levitt, in an article discussing pitch contours, stress that because of the experimental difficulties involved, they adopted a complex quantitative system involving orthogonal polynomials.

While we are sympathetic to the need for quantitative methods in speech research, it is important not to lose sight of the importance of the non-numeric methodology that is central to linguistics. Statistically oriented research may succeed to a certain extent, but its general emphasis is, in our opinion, misdirected. By peering too closely, as it were, at the data, many interesting (and, I believe, useful) facts about language are overlooked. If the pressure for an inductive, statistically verified algorithm is too great, many appealing but hard to prove statements about language will not be examined in sufficient depth.

---

[10] Cf. Nakatani, Olive, and Olive and Nakatani for other examples of this approach.

In general the problem of evidence is a hard one. The analogy is usally made between linguistics and physics, insofar as both have to abstract from the actual lowest level data in order to make any useful general statements. The analogy should not be pursued too far since at least in physics some of the conflicting influences that distort the predicted into the actual are known and can be controlled for, if not eliminated. In language study, we are very far from a knowledge of even the most rudimentary conflicts in enough detail and precision to be able to accommodate interferences with the abstract predictions. Thus rather than try to absolutely verify (statistically certain to within epsilon) some linguistic hypothesis, we must be content to demonstrate the reasonableness of the hypothesis. If we can achieve a situation in which the claims are not disproved, and are shown to be in fair accord with the data, we can incorporate the hypothesis into our system. A choice between equally "reasonable" competing hypotheses would be decided, not on an evidential basis, but in terms of the theoretical linguistic interest of the hypotheses.

We will present two instances of statistical methods applied in our speech synthesis research below: one related to the question of durational shortening as a function of word function and utterance length; the other an attempted least-squares best fit for pitch contour. The results in the first case display some statistically certain effects, as well as some that are only suggestive. The second experiment produced no usable results and was discarded as an ineffective tool for this research, being replaced by less statistically oriented methods. We feel that the suggestive results can be borne out by other, non-empirical criteria, and that this verification (in linguistic terms) should be sufficient.

I.1.2.3 Synthesis two

We have in the above two accounts a conflict between computer scientist qua theoretician and computer scientist qua empiricist[11]. The unifying fact about different computational outlooks is that all of them rely on building a 'working' model of the subject under consideration. Up to that point, and with reasonable leeway on the subject of what 'working' means, most computational researchers (possibly most reasonable people, also) would agree. Not agreed upon is the method of verifying what is a working system and the question as to what is an interesting system, whether working or not. Hopefully, views from both conflicting parties are represented in the work reported below, but the emphasis is certainly greater on the theoretician's way of looking at things.

There are only a limited number of statistically verified results reported below. We are not trying to ignore the question of verification. Our guiding principle in constructing the synthesis system has been to model our theoretical hypotheses as cleanly as possible, so that we can specify what parts of the system represent any particular hypothesis. By doing this, we make each hypothesis more verifiable even though the overall complexity of the system may make it hard to verify the accuracy of the total model, except in non-rigorous terms. For example, in a study we did (discussed further below, Section IV.1.1) on the length of our generated sentences, the overall system performance could only be loosely verified because several components and independent hypotheses were involved. The structure of our lexicon, the method of assigning parts of speech, the association of shortening factors with part of speech (which was what we really wanted to test), the modification of duration shortening due to syntactic structure, and the syntactic analysis and its effect on selecting part of speech and duration shortening were all included in the test as conducted. In addition, a hypothesis that is currently popular that there is a non-linear relationship between sentence (or phrase) length in words and sentence length in seconds was also important in assessing the results of the experiment. While the results of this test were generally reassuring with respect to the overall system performance, no precise measure of performance can be given. Finer tests along the same lines are expected to give hard data on the central hypothesis about duration shortening.

---

[11] One jargonized equivalent for empiricist is hacker. I refrain from repeating any of the common equivalences for theoretician.

At the same time that we have been trying to maintain theoretically reasonable modules as sub-components of the system, an effort has been made to develop an efficient overall algorithm for speech synthesis in terms of these modules. The modules of the speech synthesis program interact in highly limited ways. For example, the output of the syntactic analysis is a bracketed string (with individual words marked for part of speech). The bracketing indicates the surface syntactic structure (i.e. the constituent structure) of the string. This bracketed string, which does not attempt to encode, e.g. transformational history, is the basic input to the prosodic assignment component of the system. Other information, such as underlying contour to be selected or initial pitch and other similar parameterizations of the speech model are indicated separately. By constructing a synthesis system based on modular components like these, we are making claims about the linguistic systems required for human intonation, about the nature of these systems, and about the interactions between these components. While we do not claim that our program provides a neural map for the parts of the brain responsible for intonation, to the extent that our model is valid we have provided an abstract limitation on what that map might look like.

I.1.3 General synthesis

The research reported here can be designated 'research on cognitive structures related to intonation'. As a branch of linguistics this research attempts to identify human characteristics by discovering abstract and generalizable results about the way language operates. As a branch of computer science there is a need to achieve the level of specificity necessary to develop a working system for speech synthesis. From the point of view of the linguist, whether phonologist, syntactician or other specialist, the crucial question involves generalizing over specific empirical statements in a terminology that makes claims about universals of language use (and human behavior deduced from these claims). The technical bias demands a more specific set of claims and a potentially harsher set of tests. Thus while the linguist might be satisfied with a partial solution of a complex problem where there were exceptions to the general rule but where that rule captured an interesting insight into that problem, the technical person would need to actually enumerate the set of exceptions, and would probably prefer a less intuitive but more complete solution.

One example of this sort of conflict can be seen in the handling of the lexicon. There are many vagaries of the actual lexicon used for this project. These vagaries do not impinge on the abstract claims made (as a linguist). They do (critically at times) interact with the synthesis system. For example, if some of the individual words in the lexicon were recorded faster than certain others, then when a phrase is synthesized, that word will sound rushed and the synthesized sound quality will suffer. Obviously this (minor) technical problem does not negate the general theoretical basis for the system. It does create a technical problem however. Another example related to the lexicon is the question of syllable position in a word. Supposing that a given syllable in each word is assigned a prominence in the abstract system. Since the technical system which is employed for this research project[12] involves a lexicon which has no inherent notion of syllable, an approximation to that definition had to be made[13]. To the extent that the approximation was a good one, the actual model will bear a close resemblance to the abstract model and both can be judged empirically to at least a general measure. To the extent that the two models diverge, i.e. that the technical translation of the abstract model is not a good one, we do not need to deny the validity of either separately.

---

[12] Especially the LPC recording of the individual words.

In conclusion we would like to quote from Edward Sapir (1925)[14]:

> The whole aim and spirit of this paper has been to show that phonetic phenomena are not physical phenomena per se, however necessary it may be to get at the phonetic facts by way of their physical embodiment. (p. 25)

We do not intend to disregard facts about the real world (as recorded and reported in physical, phonetic, or statistical terms), but our overriding concern is with the human element which is crucial to the analysis and synthesis of human speech.

I.2 A note on terminology

We will be using the terms 'pitch', 'fundamental frequency', 'frequency' or F0, and 'tone' fairly interchangeably. The difference can be characterized as one of viewpoint more than substance, in that 'fundamental frequency', 'frequency', and F0 refer most appropriately

---

[13] Also, an approximation to an algorithm for assigning prominence to syllables for English words had to be developed.

[14] Sapir uses the word 'phonetic' where he obviously means 'phonemic' in its modern sense.

to the data viewed phonetically; 'pitch' is most appropriate for phonological descriptions, especially in reference to similarities between music and speech, and, 'tone' is the concept specific to the system that is promulgated here. 'Tone' is the least accurate term since it designates a level sub-contour which is the value of the tone, but it does have a reasonable intuitive meaning. Also, the terms 'intonation' and 'prosody' (as well as their morphological variants) will be intentionally confused, even though prosody is a more general term encompassing the pitch contour (to which the term 'intonation' should be limited) as well as rhythmic and other systemic features of speech. Some more specific terms such as 'suprasegmental' or 'autosegmental' phonology may also, from time to time, be used in place of either 'intonation' or 'prosody'. Except in cases where the difference between the terms will be confusing, they will be used in the most general construal of their meanings.

Another possible terminological confusion involves the words 'loudness', 'intensity', 'amplitude' and 'gain'. Again, differing traditions and points of view are reflected in the choice of one of these terms over any others. 'Loudness' may be the least technical choice; 'intensity' may have some psychological connotations as well as acoustic significance; 'amplitude' most directly refers to the height of some curve designating the energy of the sound, and, gain is an electronics related term which is especially relevant to the recording and reproduction of sounds. We will usually use 'loudness' in what follows, occasionally slipping into one of the jargonized terms for special effect.

Not wishing to indicate any smaller degree of confusion for terms pertaining to time, we can notice that 'length' and 'duration' can refer to the amount of time required to complete an utterance. The word 'time' itself can serve this function as well. The real confusion here is that 'length' could easily apply also to any other linearly organized and measurable quantity. For example, we could be referring to the number of successive storage locations required for the description or storage of an analyzed sound as its length. This particular use of 'length' would be linked in a general way to 'duration' but we will try to stick to the latter, since 'duration' is less likely to mislead.

We give below, Figure 1, a table of the possible conflagrations of meaning that can creep into this document.

PREFERRED	OTHER TERMS
pitch	fundamental frequency, frequency, F0, tone
loudness	intensity, amplitude, gain
duration	time, length
prosody	intonation, (supra-segmentals)

Figure 1. Preferred and confusing terms for concepts related to speech synthesis.

I.3 Some limitations of the research

The research reported herein does not cover all of the various circumstances of intonation in English. The first limitation is that most of the results reported are limited to declarative sentences.

Furthermore, the bulk of the data drawn upon (and the entire recorded lexicon) was recorded over a three year period by a single male speaker of college/post-college age. Only utterances that were read from a prepared text which the reader familiarized himself with in advance of recording were considered.

A word or two about why these limitations need not be viewed as severely limiting the generality of the linguistic claims that are made:

1) **competence-performance:** The difference between read and spontaneous text can be related to the competence-performance distinction. In a series of experiments on pauses in spontaneous speech, Goldman-Eisler (1972) found that most of the pauses reflected planning time that the speaker needed to figure out what to say next. We have not attempted to model this 'performance' feature of speech and would also be content to ignore similar performance irregularities in the pitch contour. If it were the case that speakers tended to impose pitch contours (randomly) due to memory limitations or to not planning the utterance, our model as a competence model would still be valid. In a more practical light, we are not inconvenienced by having produced a slightly better than average text reader as a result of our research.

2) **single speaker:** A valid criticism of this work

is that a good deal of it is based on the recordings of a single speaker, who may have certain idiosyncratic ways of speaking. At least one instance of an idiosyncratic speech style for the recording was noted. Two methods for avoiding the problems raised were employed. We have included data from other speakers, male and female. Additionally, a good deal of the results were finalized only after the use of listening tests, in which alternative versions of selected utterances were compared.

3) declarative sentences: We have limited most of our claims to declarative sentences. Although the material examined was largely logic and science oriented, it was not exclusively so. A variety of sentence structures were recorded and analyzed. There may well be additional factors to analyze and it is also possible that some of these may make critical differences in our proposed analysis. In a different speech register (other than the primarily read speech we have investigated) the prosodic structures might be radically different. The more general claims, developing a phonological system to describe intonation in English, rest only in part on the empirical results discussed here. Additional motivation for those claims is presented below in the

full discussion[15].

I.4 Overview of the dissertation

While some simplifying assumptions have been made in the course of this research and several areas of importance to the overall project have yet to be fully explored, the results and conclusions contained in this document represent a coherent attempt to deal with some of the most crucial issues and to provide a framework for dealing with most of the still untackled problems. "Normal" declarative sentence intonation receives the most attention here, providing the theoretical basis for a very general treatment of pitch in human language. Perturbations of normal intonation related to various kinds of perceived stress phenomena and non-declarative structures are also explored. The interrelationship of syntax, semantics and phonology in regard to intonation is examined, although a good deal of research remains to be done before we can claim total control over these processes.

Specifically, the systems that are treated below are: phonology (and phonetics), syntax, and semantics/text-view.

I.4.1 Phonology

This account is primarily concerned with the phonological relationships existing between words. We will not be concerned with the problem of word level phonology except insofar as such a

---

[15] See also Liberman (1975).

consideration aids solution of the problem of phrase, sentence and paragraph level relationships. We will discuss the autosegmental hypothesis as part of the synthesis model, as well as developments that transcend the scope of the original hypothesis. In this connection we will consider the relationship of music to speech, this relationship to be described at an abstract level.

The treatment here is largely concerned with the development of a fundamental frequency model for speech synthesis, but will also touch on metrical issues in brief. Also, while the bulk of the discussion will be related to "neutral" intonation patterns (especially declaratives), there will be application of the system developed for neutral contours to non-neutral (or stress) contours. The actualization of the proposed phonological treatment into a reasonably specified phonetic account is also provided.

I.4.1.1 More detail

It may be useful to provide the framework for the phonological account to be given. The general idea flowing through the phonology here is that speech and music are intimately related, both at the descriptive and abstract levels. At the descriptive level we can notice that both sorts of sounds can be described in terms of level tones (either musical notes or tone levels phonologically defined in terms of the phonetic fundamental frequency). At the abstract level we will assert (and try to justify) a similarity between musical analysis and our intonation analysis.

Given this ideology we can look at some concrete examples. We can notice that the relationship of tone levels for successive words in a sentence, phrase or utterance generally have a downstep relationship. Downstep means that successive tones get progressively lower by some stepping amount. We can see this in two and three word proper names, and short sentences as seen below, Figure 2.

WORD	TONE (hz)	WORD	TONE (hz)
Pamela	154	Pamela's	130
Jackson	147	dreaming	112
Richard	200	people	133
Milhous	177	cursed	128
Nixon	150	Nixon	110

Figure 2. Some two and three word proper names and sentences.

Life is not so simple, however. Further research shows that the syntactic structure of the utterance must be taken into account when assigning the intonation contour. Declarative sentences, with a neutral intonation, have a contour that can be abstractly described as (M)-H-L, optional mid tone followed by a high tone followed by a low tone. This description is actually applied recursively to each phrasal constituent to achieve an actual contour for an utterance. For example, in Figure 3 below we see a possible application of the (M)HL contour to a bracketed string (line 1). The first application of the contour (line 2) shows that the optional mid tone is not used at this level of application. The high tone has widest scope, ranging over the entire sentence with the low tone being applied at the very

```
[    [ A B ] C [ D [ E F ] ]      1
[(M) H----- H-S H-2S -------L]    2
  | [ M HL]   [(M) H---L]         3
  0                [ M HL]        4
```

Figure 3. An abstract example.

end of the sentence. Each constituent gets a downstepped high tone (written H-S or H-2S, where S would be the size of the downstep). Within each constituent the contour is again applied (lines 3 and 4). This time the tones are interpreted within the context of the overall contour so that the high tone that is assigned to B is higher than that assigned to F. When two tones get assigned to a single word, the result is a moving tone (rise or fall) on that word. While the same sort of description may be valid for word level phenomena we are not, in general, going to be concerned with this issue here.

Other sentence types have other contours describable in the same sort of autosegmental notation (more will be said about autosegmental ideas in the dissertation) and applicable in the same recursive way. Unusual (semantically or otherwise derived) stress patterns can also be described in this terminology. For example, emphatic stress can be described as the assignment of a highest high tone to the stressed element[16].

---

[16] This example requires a fuller discussion than is appropriate here. Saying 'highest high' hides a multitude of processing features. Also, I don't want to suggest that only fundamental frequency determines stress perception. Again, more about this will be said in the dissertation.

A phonetic specification of the phonological assignment takes the tone levels assigned and smooths the pitch curve between the pitch peak of one word and the lexically determined beginning (or projected beginning for sounds that begin without voicing) of the following word. The last word in the sentence, as well as phrase final words (in major phrases) are "unsmoothed" with respect to the word that follows. The last word maintains its lexical reading, while phrase final words fall more than would be predicted from the pitch at the beginning of the following word.

I.4.2 Syntax

As mentioned above, the surface syntax of an utterance affects the intonational contour that is produced in natural speech. Thus there is a section of this work devoted to the construction of a (fast) surface syntactic parser. The parser uses only limited lexical information, such as part of speech, in arriving at a constituent bracketing of an input phrase. Tradeoffs between speed, access to deep lexical information, and accuracy of parse are discussed. The importance of 'understanding' to speaking is discussed. Punctuation scanning as a component of a syntactic parser is also discussed.

I.4.2.1 The parser

The parser developed for use with the intonation synthesis project is a surface structure parser. The result of a parse of a phrase or sentence is an unlabeled, bracketed string containing the input words with an annotation of the part of speech of each word in addition to the nested bracketing. If it had been desirable, a labeled bracketing could have been produced instead of the unlabeled bracketing, but labels were not found useful as input for the phonological component.

There are several points of interest in this parser. First, there is an important question of linguistic generalizations captured in this program. This parser is fairly "dumb". That is, except in a small (less than five) number of cases it uses only the part of speech of the word to determine the parsing. While some of the complexities of necessary information could have been coded into the parts of speech, this is not the case here, where only eight different parts of speech are legal[17]. Part of the simplicity of the parser is possible since the underlying assumption is that the input strings are grammatical. Furthermore, even if they aren't grammatical, they still have to be parsed and spoken!

It can be argued that the dumbness of the parser is actually a tribute to the strength of structure preserving constraints in English. Suppose that Chomsky's most recent proposal about syntax (1977), which limits the transformational component to one structure-preserving movement rule, is essentially correct. Then the surface structure of a sentence should look almost exactly like its deep structure modulo certain alterations in the labels attached to various items and allowing that a (very strong) application of the "MOVE

---

[17] They are: N(oun), V(erb), C(opula), A(d-term), P(reposition), D(eterminer), R(educed form), X(conjunction).

SOMETHING" transformation may have moved a phrase (or a term) to certain locatable spots in the surface structure. The surface structure should be relatively easy to parse since the deep structure is supposed to be generated by a context-free system. In addition, the X-bar hypothesis allows us to confuse different names for a node or lexical label in most cases; i.e. it isn't crucial whether we think a word is a noun or a quantifier since these have nearly identical X-bar structures available to them.

The second interesting question that this parser raises is the question of the relationship of "understanding" to structural analysis. Since the parser makes no reference to the semantics of the utterance, either truth-functional or otherwise, it certainly doesn't do any understanding of the utterance. Indeed, the parser doesn't even check for case or number agreement[18]. To the extent that it produces a reasonable (to a human observer) parse, it suggests a possible definition of structural analysis in which understanding is not crucial. A weaker statement of this result might be that it negates the view that understanding is important to the structural analysis needed for speech (intonation) synthesis. Even the weaker formulation is a non-negligible result in light of much controversy between autonomous syntax and its opponents. Another reference to the issue of understanding as necessary for analysis, will be made in relation to text-structural analysis.

---

[18] The omission is deliberate, and, since even (grammatically) incorrect sentences can be spoken aloud, this is a necessary feature. Furthermore, human speakers often have incorrect structural analyses of their utterances, for example in regard to case and number agreement facts.

There are also some "technical" issues of interest in the parser such as the use of punctuation to guide a parse, and the non-recursiveness of the parsing program (achieved by using a layered iteration parsing, with similarities to the "wait and see" parser developed by Marcus (1974, 1976)).

A final note on the parser is that after a full (often successful) parsing attempt, it allows a "Total Bracketer" to work. This "Bracketer" uses a conviction that English is a right-branching language to attach (with additional levels of brackets) any un-bracketed words to bracketed constituents to the right. The theoretical status of this Bracketer will be considered.

I.4.3 Semantics and Text-view

We have created a two place "garbage bag" for dealing with other linguistic phenomena that interact with the syntax in the production of intonation contours. These two systems are separated by the scope of the phenomena involved. For our purposes, semantic phenomena are those that can be identified on a sentence by sentence basis, without recourse to the entire text being spoken. Text-view phenomena require access to whole sections of text in order to make sense.

Certain semantic relationships are identified and related to intonation synthesis. That is, given the phonological means of producing stress (or other contours), some marking scheme for denoting these relationships needs to be incorporated into a full scale system.

We are not concerned here with truth-functional semantics, but rather with functional, and lexical semantic effects. This discussion will be somewhat sketchy and less integrated into the total system than syntax.

Our use of the term "text-view" touches on pragmatics, conversational postulates, and discourse analysis. We propose (in outline) a method for selecting textual features, such as subject, topic, etc. without resorting to an in depth understanding of the text. We will show that a normative form for many texts can be used to achieve a reasonable (with respect to intonation synthesis) structural analysis. Questions of multiple sets of systems for different texts as a method of utilizing normative forms will be discussed. An appendix showing an analysis of several lessons taken from a current logic curriculum is included.

I.4.3.1 Semantics

The semantics we are concerned with for this project is simply a way of designating those interferences with the normal (neutral) assignment of intonation contour as dictated by the syntactic structure. In one sense the semantics could affect a re-bracketing of the text at hand, but it is much more realistic to want to mark the items to be stressed or de-stressed as such. Currently, text to be synthesized can be explicitly marked to designate stressing of words. The function of the semantics (and also of the text-view component) is to narrow the class of cases for which a human has to mark the case explicitly. While normal focus and topic/comment intonations are built into the syntax, any deviation from the normal status should be picked up automatically and marked explicitly.

I.4.3.2 Text-view

An immediate structure analysis should be possible within the realm of extended text. By extended text (henceforth, simply: text) here, we mean any collection of utterances, written or spoken in some unified manner. In our sense, a chapter from a book, a journal article, or a lecture given are all examples of text. We have chosen logic lessons for this example because of the clarity of the structure of the lessons. While we consider the structures discussed below to be valid for generalized text analysis, motivating them will be easier in this domain[19].

We have identified a relationship betwen constituent text structure and relative sentence durations which encourages the idea of constructing a series of recursively nested structures by which to analyze text into constituents. The constituent analysis being proposed here is not an attempt at full understanding of text. Our goal here is to account for as much of the purely linguistic structure as possible without having any explicit real world knowledge included in the model. This analysis will have a connection to prosodic generation as well as possible application as a knowledge source in a more general understanding system. By constraining the class of data we are dealing with, as well as the expected results of the analysis, we hope to be able to account for larger sections of text in a useful way.

---

[19] It is possible that different organizational systems of constituents will be required for different types of text.

I.5 The computer environment of this research

The computer system at IMSSS consists of a central computer with its associated random-access core memory, a drum used for swapping (a timesharing system feature), magnetic disks used for information storage, magnetic tapes used for off-line information storage, and user terminals. The computer is a Digital Equipment Corporation KI10 running the TENEX timesharing system. Core memory consists of a quarter of a million 36-bit words with an access time of 300 nanoseconds and there is on-line disk storage of over one hundred million (36-bit) words[20]. In addition, there is one high-quality audio recording/playback booth, a specially built speech synthesizer, described below, and user playback sets (generally headphones) attached to many of the terminals.

A block diagram of the complete hardware system can be seen in Figure 4.

I.5.1 The nature of the lexicon

A lexicon consisting of several thousand individual words has been recorded, and each word has been analyzed into linear predictive coefficients for efficient storage and prosodic manipulation. Twelve

---

[20] Each computer word can store five characters, three twelve bit numbers or one thirty-six bit number.

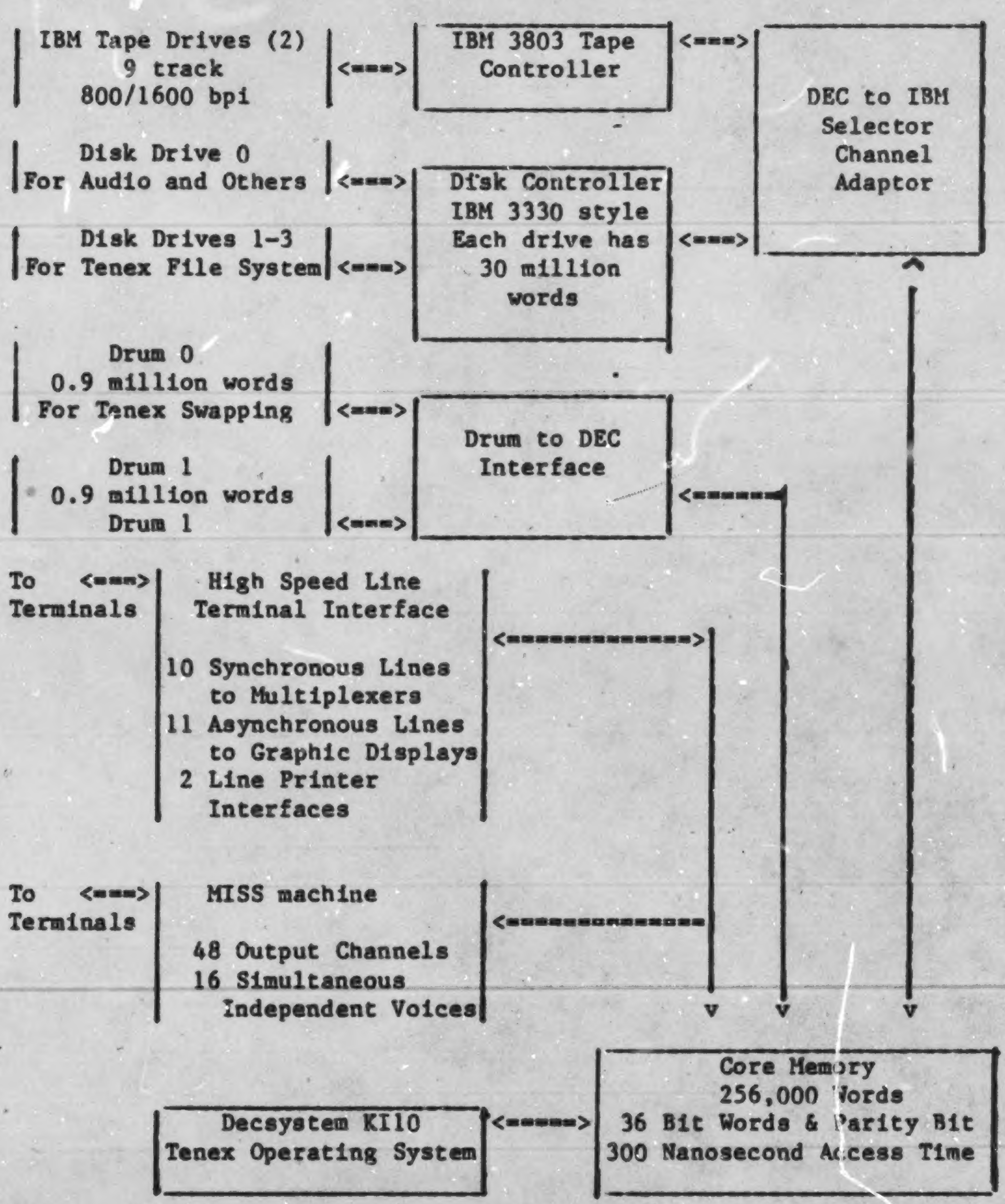

IBM Tape Drives (2)
9 track
800/1600 bpi
IBM 3803 Tape Controller
DEC to IBM Selector Channel Adaptor
Disk Drive 0
For Audio and Others
Disk Controller
IBM 3330 style
Each drive has 30 million words
Disk Drives 1-3
For Tenex File System
Drum 0
0.9 million words
For Tenex Swapping
Drum to DEC Interface
Drum 1
0.9 million words
Drum 1
To Terminals
High Speed Line Terminal Interface
10 Synchronous Lines to Multiplexers
11 Asynchronous Lines to Graphic Displays
2 Line Printer Interfaces
To Terminals
MISS machine
48 Output Channels
16 Simultaneous Independent Voices
Core Memory
256,000 Words
36 Bit Words & Parity Bit
300 Nanosecond Access Time
Decsystem KI10
Tenex Operating System

Figure 4. Block diagram of the computer system.

coefficients are used to predict the shape of the vocal tract for each frame (small unit) of the sound. In addition, the duration, the pitch period and three loudness parameters are stored for each frame. The algorithms in the sentential synthesis system only modify the duration, pitch and loudness parameters of the stored words. It has been suggested that smoother word transitions would result from working with the other predictor coefficients as well[21]. The MISS synthesizer is specifically designed to efficiently produce speech from these linear predictive coefficients. It is also capable of calculating changes to the duration, pitch and loudness parameters in real time.

Each word in the lexicon is treated as a unique entity, unrelated to any other word in the lexicon. All morphology, segmental phonology and morphophonemics have been eliminated from the lexicon and the prosodic system. Even such regular morphologically related forms as singular and plural of a word are placed in independent entries in the lexicon. For example, the words "book" and "books" have entirely separate listings and representations in the lexicon. They could have been recorded at times well separated from each other, and we are making no claims as to any regularities between such morphologically related words[22]. We do expect that there is a certain amount of regularity for the entire lexicon. If we consider the parameters which are stored for each of the lexical items, the regularities we expect are as follows:

---

[21] More technical descriptions of the linear predictive coding, analysis and storage can be found in Sanders and Benbassat (in preparation) and in references given there.

[22] Not that there are no morphological regularities in English. The utilization of these regularities goes beyond the scope of this dissertation. See, for example, Hunnicutt (1976).

1) We modify the fundamental frequency of the word by multiplication of part of the contour by some constant and imposition of a linear contour replacing part of the original contour[23]. The regularities which the system depends on are that there be a reasonable pitch peak at the 'right' place in the recorded word, and, that the recorded pitch curve be realistic on at least the section of the contour leading to the pitch peak.

2) The loudness curve for the recorded word is normalized so that all the words are at approximately the same loudness. The regularity assumed is that the loudness contour for any word is reasonable. In fact, a reasonably important problem for speech synthesis, but not dealt with in this dissertation, is the question whether the loudness contour for a lexical item is similar to its contour in a longer utterance.

3) We modify word durations by multiplying the words by a fraction of their lexical length[24]. The

---

[23] Modification of the pitch contour is discussed in greater depth later, Section III.5.

[24] Some possibilities for duration modification are discussed below, Chapter IV.

assumed regularity here is that the speaker was consistent in lexical duration as he recorded. That is, we assume that a word that is inherently longer in lexical form, will be appropriately long as recorded. Studies by Gaitenby, et al. have shown that while there may be large differences between different speakers (and dialect areas) in regard to speech speed, each speaker is fairly consistent in his or her own utterance speeds. In practice, this assumption of regularity has been born out. There have been some irregularities which can be noticed when listening by the seeming rapidity which they posess when used in synthesized sentences. Further studies using phonemic spellings could be used to predict approximate lexical length for the words in the lexicon and to designate those that are not normal with respect to the rest of the lexicon.

4) The other vocal tract predictor parameters are left unmodified by the synthesis procedures. Thus we have no inter-word smoothing of the predictor coefficients (and thus no smoothing of the higher formants). We also need to assume that the predictors are reasonably accurate when the pitch of the utterance is modified. Since the parameters were determined for a given word, partly as a function of

its pitch contour, when we modify the pitches of the contour the predictors are no longer quite accurate. We have found that by staying reasonably close to the pitches of the original contour, the vocal tract predictors remain fairly accurate and produce a good quality sound on output.

Along with the predictor coefficients, pitch and loudness contours stored for each word, the lexicon contains additional information that has been found to be useful for doing prosodic manipulations in real time. The pitch at the beginning of the sound if it is voiced, and the maximum pitch of the sound (in hertz) comprise part of the lexical entry, along with the part of speech of that word. Additionally each word is marked for the location of its maximum pitch.

A representation of the lexicon is given below in Figure 5.

In the figure, we show that there are two parts to the audio lexicon: a quick lookup (hashed) table for storing the information about the word that is required for the contour assignments, and the main storage area, containing the actual phonetic specifications of the words. The phonetic store is used by the synthesis hardware to generate the sounds, subject to the phonetic controls specified for the contour assignments in the command string.

```
|----In a quick lookup table----||Pointed to by a quick lookup pointer|

WORDi PART-OF-SPEECH  PEAK  FIRST   FRAME1 FRAME2 ... FRAMEj ... FRAMEn
      (2 are allowed) PITCH PITCH   LPC1     ...  ...(marked)...   ...
                                    LPC2     ...  ...  (for) ...   ...
                                     ...     ...  ... (peak) ...   ...
                                    LPC12    ...  ... (pitch)...   ...
                                    DURATION ...  ...   ...    ...   ...
                                    PITCH    ...  ...   ...    ...   ...
                                    GAIN1    ...  ...   ...    ...   ...
                                    GAIN2    ...  ...   ...    ...   ...
                                    GAIN3    ...  ...   ...    ...   ...

       Used to assign phonetic      Sent to hardware synthesizer to be
        specification controls.     modified as part of the word
                                    concatenation procedure, subject
                                    to phonetic specification controls.
```

Figure 5. Schematic representation of the audio lexicon.

Chapter II

English as a tone language

The system we have developed for speech synthesis has been largely concerned with the prediction of pitch contours for English sentences and phrases. We present below a description of the phonological framework which guides research relating to the synthesis system. We will construct and theoretically justify our system in this chapter, while in the following chapter, we will present the experimental justification for this approach.

II.1 Preliminary discussion

Let us begin description of the phonological framework by stipulating the constructs that will be involved. Later, we will discuss the processes that utilize these constructs.

The fundamental construct for our phonological system is the tone group. A tone group is composed of a sequence of abstract tones. While there is no a priori limit to the number of tones in a tone group, we expect that there will be an inverse relationship between length (number of tones in the group) and popularity. The more tones required for the description of a tone group, the less frequently we expect it to be used by a language. This expectation, combined with a principle that there is a unique description of each tone group, such that no two adjacent abstract tones are the same[1], provides us with a metric for judging the reasonableness of different descriptions of the tonal system of a language when given in 'tone group' terms.

At this point in the discussion, we see that the 'contradiction' tone contour given in Liberman and Sag (1974) and discussed above, Section I.1.2.1 can be accommodated in our system of tone groups. The complex contour given there has been reduced in Liberman (1975) to a simpler form: as a boundary high tone followed by a (in his terms) mid-low and then a high and a low. If we equate the mid-low tone with our mid tone, then we see that the contour is simply a 'declarative fall' contour with an an adjoined initial boundary tone. While this combined contour is somewhat more complex than we might otherwise want to allow in tone group system, the contour is certainly marked as unusual, and, it is not clearly disqualified (for its limited pragmatic effects) from being one of the language's tone groups.

---

[1] See below, Section II.3.2.

Tones which may appear in a tone group are selected from a language specific repertoire of tones although the set of universally possible tones is limited to a small number (approximately five or seven). These tones are similar in purpose and function to autosegmental tones (Goldsmith 1976). In addition to the abstract tones which compose the tone group, we will need a larger repertoire of concrete tones, which describe the result of applying various processes to the abstract tones of the tone group.

We will also want to allow (in our phonology) the concept of boundary tone which is an abstract tone in a tone group with a unique elaboration. The regular elaboration processes will not apply to the boundary tone. We will not actually use the boundary tones for any of the discussions here, since the details have not been worked out, but we wish to include it in our formal repertoire for completeness.

We will use an intuitive concept of word. We compile a lexicon consisting of all of the words for which we have a phonetic specification. Closer to the phonetic level, we define three pitch points which are associated with each word in the lexicon: the initial pitch, the peak pitch, and the final pitch. Each pitch point has a value and a location. The initial pitch is defined as the pitch (fundamental frequency) of the first segment if the segment is voiced, and zero if it is not. The peak pitch is always strictly positive (never zero) and is the maximum for the lexical (isolation) pitch contour of the word. If there is more than one pitch maximum in a word, we will choose the last one. Obviously, the value of the peak pitch is indifferent to the choice of maximum, but the choice of location may be significant for some of the processes we will define below. The final pitch is the pitch of the last voiced segment in the word. The last voiced segment need not be the last segment in a word. Note that, in theory, either the initial or final pitch may have the same value and/or location as the peak pitch, although in lexical form in English it is very unusual for the final pitch to be the maximum. This fact is not inherent in the system we are constructing, and we will see that in some contours the final pitch may equal or exceed the value of the pitch at the location of the lexical maximum.

Finally, we need syntactic and semantic constructs that are motivated independently from phonological considerations. The constructs that we will employ are constituent structure, part of speech, and stress. Constituent structure is the organization of a linear string of words into a tree structure, represented by a bracketing of the linear string. The part of speech of a word in a given syntactic context is the same as is the standard grammatical notion bearing the same name[2]. By stress we mean the indication of non-neutral semantic content in the context defined by the string (as well as preceding and succeeding strings). We allow several degrees of positive stress, ordered from most stress to least, and, (at least) one degree of un-stress or stress reduction.

---

[2] The actual implementation of this, as well as other constructs and processes, may differ from the abstract definition. For example, the part of speech is currently implemented as being independent of the syntactic context. In fact, it is used in the generation of a constituent bracketing.

While it is conceivable that further research into syntactic and semantic interactions with intonation will result in the addition of other constructs to this system, there are some possible additions which we do not expect. One of these is the introduction of phrase labels, such as 'noun phrase' or 'verb phrase' since the current results indicate that they are irrelevant; i.e., the same processes occur in differently labeled phrases.

II.1.1 An example

Let us assume that the tone group in question is the "declarative fall", which we can write out as the sequence of tones: "(M) H L". Let us take a sample sentence (string of words): "John saw the dog." We will give it the following constituent structure:

[John [saw [the dog]]] .

For each word in the sentence we can give a part of speech and three pitch value variables, as in Figure 6. The absence of any explicit stress marking implies neutral stress. It will turn out that we can use the peak pitch variable to also represent the concrete tone for the word.

We define a hierarchy of concrete tones to be the tree declarative fall

(M) H L

[John	[saw	[the	dog]]] .	
N	V	D	N	part of speech
IP1	IP2	IP3	IP4	initial pitch
PP1	PP2	PP3	PP4	peak pitch
FP1	FP2	FP3	FP4	final pitch

Figure 6. Description of a pitch contour.

structure formed by choosing the maximum of the peak pitches for each word in a constituent as the peak pitch for that constituent. The hierarchy for the above example is represented below in Figure 7.

[John	[saw	[the	dog]]] .	
PP1	PP2	PP3	PP4	
PP1	PP2	PPj		(j = 3 or 4)
PP1	PPk			(k = 2 or j)
PPm				(m = 1 or k)

Figure 7. A hierarchy of concrete tones.

While the tree represented above is binary, we also allow multiple branching hierarchical structures.

We have so far presented the picture of an abstract description of the pitch contour of a sentence. To broaden the picture to include other intonational variables is straightforward. We introduce duration and loudness variables for each word. These variables are functions of the part of speech, syntactic context and semantic context (presence and type of stress) in a way similar to the pitch variables, but are also functions of the lexical duration and loudness, respectively (as is the peak pitch, for example).

II.1.2 Continuation of the description

The next step in the description of the phonological system is to detail the processes that flesh out the abstract description. We need processes which assign values to the above variables, and then others which produce a phonetic specification of the intonation contours. The processes which we make available are: elaboration of the tone group; mapping of tones onto words; assignment of pitch values based on word tones, and, smoothing of the contour between the different pitch assignments.

II.1.2.1 Elaboration

A tone group is elaborated by means of universal devices to generate the tonal contour of a string. While one tone group is sufficient to describe a phrase (or sentence) at an abstract level, the tonal contour of that phrase will usually be more complicated. The tonal contour we are referring to here is equivalent to the autosegmental "tonological" specification of a sentence as used by Goldsmith (1976). The elaboration of the tone group is responsible for generating the hierarchy of concrete pitches which was illustrated above.

II.1.2.2 Mapping

The device we call mapping is indistinguishable from the mapping of tones to syllables as done in autosegmental phonology (Goldsmith 1976), with the exception that where autosegmental phonology maps its tones onto syllables, we map the tones onto words. We will describe the autosegmental system in greater depth, below. In our system, mapping plays only a very limited role.

II.1.2.3 Assignment

Assignment of pitch values to the three variables for each word is a relatively straightforward procedure. The peak pitch of a word is assigned the value of the concrete tone associated with the word which resulted from mapping tones onto the words of the sentence. We can think of the assignment of the peak pitch value as a quasi-musical transposition of the pitch contour (melody) of the word, from its lexical "key" to its assigned "key".

The initial pitch assigned to a word is the transposed value of the lexical initial pitch, in parallel to the peak pitch. It is possible[3] to assign the peak pitch (both in value and location) to the initial pitch, i.e., the peak pitch can be forced to the beginning of voicing in the word.

The final pitch is normally assigned with respect to the transposed value of the initial pitch of the following word and the assigned value of the peak pitch of the current word. If the initial pitch of the following word is non-zero then the final pitch is assigned so as to be equal to it. If the following initial pitch is zero then the final pitch is defined simply in relation to the peak pitch of the current word. The final pitch can also be assigned as a result of having multiple tones mapped to a single word, as in the case of a relative high being mapped to the last word in a sentence which also has a low tone mapped to it, e.g., from being the last word in the sentence. In a case like that, the final pitch will reflect the second (in the example: the low) tone. The system developed here does not allow for more than two tones to be assigned to a single word unit. This is because our analysis is prevented, a priori, from applying too deeply to word structure. It is reasonable to think of applying our analysis to a syllabic system, but then the modes of elaboration will not be completely similar in the two systems. We discuss the issue of words vs. syllables further below.

---

[3] Experiments testing the desirability of the following are currently in progress.

II.1.2.4 Smoothing

Finally, we define a smoothing procedure which produces the pitch contour for an entire word based on interpolating pitches between the initial, peak and final pitches according to some smooth function. For example, the function might be:

> For the segment from the initial to the peak pitch, transpose the pitches lexically associated with those segments. For the pitches between the peak and final pitches, interpolate linearly using the assigned values to determine the slope and intercept of the line.

Other functions which could employ more complicated formulae might be used instead[4].

---

[4] See Section III.5.2 for more discussion.

II.1.3 Continuation of the example

To see how these different processes interact, let us examine further the example we started above, using "John saw the dog." The first process to affect the structure (shown above in Figure 6) is elaboration of the tone group. The elaboration of (M) H L in this case will yield the tonological specification: H-H'-M''-H''-L''-L'-L. The better way to visualize this specification is in terms of the concrete tonal hierarchy, Figure 8. The primed tone levels are indication of a step away from the "un-elaborated" or unprimed level bearing the same level name[5][6]. Thus H' is one step away from the "un-elaborated" tone level for H and therefore it is lower in pitch than H and H'' is two steps away from some "un-elaborated" tone level. Another explanation of the meaning of the primed levels is that they are to be interpreted as bearing their leveled tone relationship to other tones of the same "primeness" as the overall meaning of that number of primes is a function of the governing node in the hierarchy.

```
[ John  [saw [the dog] ] ] .
    H     H' M'' H''L''L' L
    H     H'     H''   L' L
    H     H'           L' L
    H                     L
          (M) H L
```

Figure 8. An elaboration hierarchy for a tone group.

---

[5] The direction of the steps is towards a central tone. Thus high tones downstep, low tones upstep, and mid tones which are closer to high tones than to low tones downstep.

[6] This statement is not obviously true (or necessary). It does reflect the system that has been developed.

Mapping takes the 'autosegmental' description and maps it onto the individual words of the sentence. In our example, the mapping is a one-to-one map for most of the tones and words. The only special part of the map is the mapping of multiple tones onto a single word. In our example, the word dog is mapped a H'' by the straightforward part of the mapping. When we try to assign all of the various L's we have run out of words! Our formalism allows us to associate at most two tones with a single word. Further, we constrain our primed symbols to remain within the scope of one syntactic constituent. We can assign the L'' to dog, but then we would need someplace to put the other L's. What we will do instead is to merge the three L's into a single tone and assign that as the second tone on dog. In effect, our convention to merge these tones is the embodiment of several tone spreading rules proposed in Leben (1976).

Once we have mapped the tones onto the words, we can assign the actual values of the different pitch variables for each word. We summarize the assignments below in Figure 9.

[John [saw [the dog]]] .

H	H'	M''	H''	L	tone contour
i1	0	i3	i4		initial pitch
h	h-s	m''	h-s'		peak pitch
f(h)	i3	i4	1		final pitch

Figure 9. Description of a pitch contour.

The peak pitches are assigned by a translation of the concrete tones into actual pitch values. The s and s′ are the step sizes relating to the two primed levels. We have abbreviated the assigned mid tone by m″ when we could have written it as m-s′. The initial pitches are simply the transposition of the lexical initial pitches[7] to the key determined from the peak pitch. The final pitch of the first word is a function of the peak pitch of that word; the final pitches of the next two words are the initial pitches of the words following them, and, the final pitch of the last word is the low pitch assigned as the second tone to the word dog. If the sentence had been more complicated, other words would have had second tones assigned to them as well which would determine the final pitch for those words.

The smoothing procedure is straightforward and results in the structure depicted below (Figure 10).

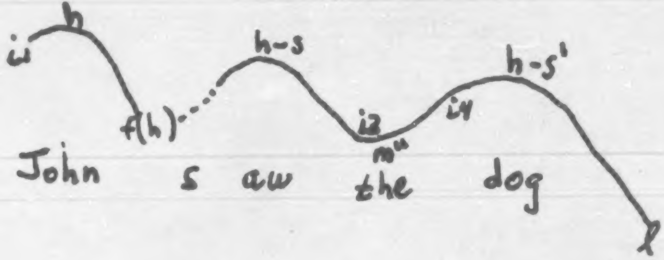


Figure 10. Description of a pitch contour.

II.2 The autosegmental hypothesis

---

[7] There is a more complicated assignment procedure that would change this assignment.

A recent attempt to describe the pitch component of tone languages has developed a formalism that also is useful for describing intonation in non-tone languages. This system develops idea that the segmentation of the tonological, or pitch, component of a tone language utterance is performed independently from the segmentation of the other phonetic/phonological components of the utterance. There are several reasons why this is a reasonable (and preferable) analysis in the face of a null-hypothesis which would say that tone is simply another segmental or vowel feature, exactly like the feature [high], or perhaps only a function of rules which operate on segmental features. First, there is an analysis of contoured tones, e.g., rising tone, rise/fall, etc., in terms of level tones, which simplifies the analysis. For example[8], in Igbo, an African tone language, there is a copying of tones from the verb to the subject which precedes the verb. In monosyllabic subjects, the tone of the subject takes precedence and is the only tone realized, but in polysyllables the tone of the second syllable will be level if it agrees with the tone on the verb, but it will be contoured if it does not. This indicates that the copying of a tone from the verb can result in either a level or contour tone. Yet the tone itself is a level tone (we are not copying contoured tones in this instance). To illustrate:

a.	o	ci	akhwa	'He was carrying some eggs.'
	H	L	L H	
	(sbj)	(vb)	(obj)	

---

[8] This example is from Goldsmith (1976).

b. Eze ci akhwa 'The Chief was carrying some eggs.'
H L L L H
(sbj) (vb) (obj)

c. Ekwe ci akhwa 'Ekwe was carrying some eggs.'
H F L L H
(sbj) (vb) (obj)

The H and L are familiar, the F stands for a falling tone. The point is that in (a) where there is a monosyllable with a tone, no tone transfers are effective, whereas in (c) the effect of copying the tone from word to word results in a contoured tone being spoken on the copied-to word. Still in (b), where (according to data not shown) the tones on the first word are the same as the tone on the verb the subject retains its level tones.

Other good analyses generated include cases where the segment (or syllable) normally bearing the tone in question is deleted, yet the tone remains in the phonology. While other analyses could also handle the same data (with the appropriate rule twisting), the autosegmental hypothesis provides a reasonable framework for these studies.

Two important ideas are part of the autosegmental hypothesis. The first is the idea of association, which was briefly mentioned above. This idea is crucial to autosegmentals. It says that given an autosegmental string and its companion segmental string, we can associate the two strings to produce the utterance and its pitch contour which are simultaneous in speech. The usual form of association has t[illegible] conditions:

1) All tones are associated with at least one

vowel[9], and, all vowels are associated with at least one tone.

2) The lines which indicate associations of tones with segments cannot cross each other.

The effect of condition 1 is that you can get more than one tone on a segment and more than one segment onto the abstract tone. Condition 2 says that the tones have to associated sequentially with respect to the segments.

The second important idea from autosegmental theory is the use of a distinguished tone element and the use of boundary tones. A distinguished tone element in an autosegmental string is assigned before the rest of the autosegmental string, whereas the boundary tone is assigned after the rest of the autosegmental string. We mentioned above that we wish to include boundary tones as possible elements of our phonological system, even though we are not utilizing them at present. They conceivably would play a role in comma intonation which seems (at the moment) to be a function of the boundary represented by the comma. The idea of distinguished element is employed in the selection of peak pitch for the lexical word. That peak pitch acts, in general, like a distinguished point in the autosegmental string. Otherwise, our view of distinguished elements is a rather skeptical one, prefering to rely on the modes of elaboration which we can motivate.

---

[9] Feel free to read 'syllable' for 'vowel' or 'segment' wherever it makes sense to do so.

In Liberman's use of autosegmental notation, the mapping from the tone string to the segments does not specify all the tones for all of the segments. He introduces some low level phonetic rules or chance to account for their tone contours. Leben's version of autosegments also allows incomplete specification of the segmental tones which are mapped. His theory inserts mid tones for unmapped tones. In addition, there are some fairly powerful rules in Leben's system which contravene some of the spirit, if not the formalism of the autosegmental system.

In our system, the autosegmental sub-component is primarily used as a notation for describing the effect of elaborating a tone group melody. The individual tones are similar to the notes of a melody. They can be transposed, or otherwise 'transformed' by certain universal[10] mechanisms.

II.3 The correspondence of tone languages and intonation

Tone languages are most profitably viewed as having a suprasegmental melody which becomes associated with the segmental phonological elements; we can adopt the same viewpoint for prosodic phenomena. A melody consists of autosegmental tones (autonomous from the phonological segments) selected from the tonal repertoire of the language. Since the melody is independent of the segments, there can be more tones than segments in an utterance or vice versa. Generally, there are two mechanisms that associate these tonal melodies with the

---

[10] cognitively universal, i.e., not language specific

segmental string. One mechanism involves a notion of "prominence": each melody has a single marked tone which is associated with the most prominent segmental syllable. General rules then account for the association of the segmental syllables with the rest of the melody. The other mechanism which associates melodies and segments specifies that the (initial and/or final) boundaries of both the segmental and autosegmental string must match. Again, the general rules of tone specification account for the association of the remainder of the melody and the remaining syllables. Different languages conceivably could use either or both methods of association. There is evidence that English is basically a prominence language; however, certain aspects of its prosody depend on the use of boundaries.

II.3.1 Tone levels

One basic theoretical prerequisite for using the autosegmental theory is positing the existence of tone levels in opposition to continuous melodies. While observed pitch contours for words and sentences rarely display level tones for any appreciable length (except while being sung), phonologically they can be described and handled as if they were at the level of the highest frequency of the contour. Most of the analysis presented in the rest of this paper talks in terms of tone levels.

II.3.2 Evidence for tone levels and such in English intonation

Let us assume that English can be analyzed as a tone language with respect to its intonation system[11]. Most declarative sentences in English have a pattern (with respect to major constituents) that has been described in Leben (1976) as

$$(M)\ H\ H{-}1\ H{-}2\ \ldots\ H{-}n\ L.$$

An optional mid tone (the terms mid, high and low, M,H,L, are phonemic values rather than actual phonetic ones) is followed by successively lower high tones which are followed by an actual low tone. In terms of our analysis, the initial mid tone corresponds to our syntactically conditioned upstep; the successive high tones are the downsteps; the final low tone is realized as the long (lexical) fall on the last word of the sentence. We would like to point out, in passing, that according to our findings presented below, the mid tone is defined relative to the position of the high tone and not as some absolute frequency (or frequency range). If we assume that the lowering of the high tones is a phonetic (perhaps universal) condition on the realization of successive tones[12], then we can abbreviate our formula above to:

$$(M)\ H_0\ L,$$

where the H_0 indicates any number of successive high tones. If we adopt a suitable convention for assigning various prominences to designated words of each sentence and relating that to the assignment of the pitches of the contour, we might eliminate the subscript from the high tone in the formula above and simply designate the method of correspondence between the intonation contour and sentential prominence[13]. Thus, our most basic "discovery" about declarative intonation in English is that it has a pattern that looks like

---

[11] The actual experimental evidence will be presented in the following chapter.

[12] See Section II.4.

(M) H L.

The complex intonation pattern for English declarative intonation that we started with has been reduced to a simpler one which is quite a familiar one from studies of tone languages and other intonational systems[14].

We should note that the step of reducing the pattern with repetition to a pattern without repetition is not a totally obvious one. Leben (1976) justifies the procedure as allowing the imposition of two restrictions on the autosegmental hypothesis (from Leben 1976, p.90):

> 1) Underlying tonal contours are independent of the nature of the segments with which they will become associated.
>
> 2) Autosegmental sequences express only transitions between unlike tones.

---

[13] A fuller discussion of this entire topic is carried out in Leben (1976).

[14] Cf. Goldsmith (1974) for a review of different tone patterns.

Goldsmith (1976) takes issue with Leben's restrictions on autosegments and argues for the possibility of tone repetitions in an autosegmental account. He argues that an account of the Tiv verb conjugations, in particular, becomes simpler if repeated tones are allowed. Whether the particular example he chose bears out the difference or not, the key point is that a reduction of the description of an intonation melody, using autosegmental terms, to a simple tone pattern is not a necessary (though it is a desirable) part of autosegmental theory.

The presence in our formalism of an initial optional tone seems to have a good empirical backing since, in addition to the declarative fall melody we also have the vocative chant, discussed extensively in Liberman (1975) which must be described as optional low tone followed by a high tone followed by a mid tone, i.e., (L) H M. Thus it is not simply a mid tone at the beginning of a contour which can be optional in that contour. Rather, any tone in initial position in a contour can be specified as an optional tone. It seems then that the simple explanation which might be offered that the optionality of the mid tone is a function of vocal cord warm-up is defeated by the other possible optional tones from different contours.

Our circle has been completed! We have provided the link back to tone languages. Based on our assumption that English prosody could be analyzed in the same terms as are used for tone languages, we have shown that at least this tone pattern employed for English prosody is one that is used by many other languages, i.e., that English prosody should be analyzed in the same terms as are used for tone languages.

II.3.3 Why words

An important question that could be asked of this project is: "Why does the system depend on word relationships rather than syllabic ones?" Indeed, syllabic information could quite probably be as easily available as word level information. It would be useful (and may be necessary) for a further extension of this project to have such information for word concatenation. The syllabic information that would be desirable, would consist of the number of syllables in a word, the peak and initial frequencies for each syllable, and the starting and ending positions within the word for each syllable. Allowing ten as a reasonable (generous) number of syllables for a word, the necessary storage is only forty-one small integers for each word, not a very great strain.

If, as we claim similar tonal phenomena are taking place inside of words as well as between words, how can we ignore this information. We have claimed that our conception of deriving the tone level as the peak fundamental frequency of a word or constituent is correct. And words and phrases are similar in having downstepping relationships between their peaks (in general). We could have imposed a syllabic pitch contour on words in generated sentences. For example, when we linearized the fall (or rise) between the peak pitch of one word and the beginning of the following word, we could have made explicit use of the peak pitches in all of the post-major peak syllables in the word instead of giving up the syllabic pitch contours from the lexical form of the word.

The basic reason not to include syllabic information in the account given here is that phrase level phenomena are more predictable than word level phenomena. While many word level phonological features can be explicated quite clearly and are very regular and rule abiding, many (if not most) are lexicalized, historical remnants, islands from sound changes, etc. and as such are exceptions to any reasonable rule system. Phrase phenomena, on the other hand, are of necessity regular with the exception of relatively few idiomatic phrases which could be considered extended lexical items in any case. The difference is also reflected in the fact that while phrases are, by and large, generated 'creatively', words are (again, by and large) simply drawn from a lexicon[15].

It is conceivable that boundary tone effects will need to be handled differently, with respect to the word/syllable distinction. Thus comma intonation could be considered a perturbation on the normal prominence contour. Then these boundary phenomena would operate to change the appearance of the lexical material[16].

II.3.4 Tone levels vs specific contour

We are discussing how our autosegmental tones become realized as a complete contour through application of general processes. It is our contention that the approach employing several levels of tones in the analysis is more fruitful than one which attempts to specify the entire contour using only one level of analysis. Here we are actually arguing against two different possibilities. One is the idea that smooth polynomials should be constructed to account for intonation contours, and the other, that no pitch contours found in a language (perhaps with some idealizations) are related to each other. With regard to the first idea, we have already argued above that it is linguistically superficial, it does not tell us interesting facts about people using language, and that it cannot easily handle the complex cases which we have shown that our proposals allow us to analyze so simply.

---

[15] A graphic example avoiding "Colorless green ideas sleep furiously" is that, insofar as possible, all the words appearing in this dissertation are already available in the lexicons of my readers, yet (hopefully) all of the sentences are brand new.

[16] This is an extremely tentative suggestion.

As regards the second idea, we need to argue that our simple autosegmental formulas (as for example, the one given above) can accommodate themselves to what would otherwise have to be claimed to be independent pitch contours. In addition to the examples of complex pitch contours given in this paper, we refer the reader to Leben (1976) for a fuller substantiation of the position taken here. Of course we are not claiming that there is only one pitch contour in English. Our system of autosegmental tones allows us to assign a large, though constrained, set of tone contours for various English utterances[17]. Still, on our analysis, English prosody is much more unified than it would be according to the alternative views cited above.

---

[17] cf. Liberman (1975).

II.4 Tone groups

We would like now to turn on the very theory within which we have been so successfully working. Our complaint with the autosegmental hypothesis is not that it is insufficient or unsuited for describing English intonation, but rather that it is not restricted enough to explain why people can use intonation properly.

Our notation here suggests that there are still more possible contours available to the language than it actually ever uses. The autosegmental hypothesis gives no reason for the specific choices of intonation patterns which a language like English makes (or could make). The autosegmental hypothesis also allows us to formulate intonation contours which seem to never occur in natural language, much less in English. Our notation so far makes the choices a language makes seem somewhat arbitrary and accidental, yet linguists have maintained that language is not capricious.

Note that the version of the autosegmental hypothesis suggested in Leben (1976), discussed above in Section II.3.2 with respect to Goldsmith's view of autosegmental theory, attempts to put some constraints on the arbitrariness of autosegmental prosodic patterns. We feel that what is required is a fairly drastic revision of the basic notions implicit in the formalism rather than simply constraining them. The constraints proposed by Leben are similar in form, function and spirit to constraints that might be put on segmental rules. Conceivably, an attempt to also organize segmental features on a prosodic basis might employ similar constructs. We feel that pitch (or tone) and rhythmic phenomena are crucially related to musical constructions. In our view, music represents a refinement of human speech. There is no logical necessity for the similarity between music and pitch (and rhythm) to imply a similarity between music and other speech phenomena (such as segmental relationships), any more than it need imply a similarity between other art forms (e.g., graphic arts) and pitch or rhythm.

If we regard the tonemic inventories of the world's languages as consisting, not of tones, but of tone groups, we discover that we can restrict the vocabulary of prosodic contours very naturally. Essentially, we wish to argue that an autosegmental sequence, such as (M) H L, is only a notation for one tone group (which we could give some name: "declarative fall"). While in autosegmental terms there are an infinite number of possible melodies (viewed as combinations of tones), in terms of tone groups (as entities in their own right) it is entirely reasonable to assume that there are only a finite number of distinct ones within the human experience.

Since we have emphasized the close similarity of speech and music it is worthwhile to discuss this matter further. Admittedly, no similarity, no matter how well motivated, can "prove" that tone groups are better than other analytical constructs, but it can indicate that pursuing the similarity is a reasonable method for placing constraints on the analysis of tone phenomena. Besides showing that we have developed reasonable constraints, we will produce some evidence that the constraints implicit in using tone groups are useful.

II.4.1 Music and speech

What is music? and why should it have any relationship to human speech? Is music simply another reflection of the general cognitive facilities that may be thought to underlie all human creation? Why should we think that speech (especially the actual production of speech) is particularly creative or artistic? What can music (musical analysis) tell us about speech (speech analysis and synthesis)?

These are questions that need to be addressed at this point. In brief, we would claim that music is structured timbre, pitch, duration and volume. Speech is also structured timbre, pitch, duration and volume. They are not identical, but they are very similar. It is reasonable to guess that the ontogeny of music is a development from speech to song to instruments producing non-speech-like sounds (and full circle to speech-like sounds again). Developmentally, then, it would seem reasonable to say that the structuring of music is a refinement or bringing to a conscious level of the structures inherent in speech. While we do not always exercise conscious control over the vocal tract components which produce speech, that does not mean that there is no artistic creativity associated with the process. Certainly, there are people (and times) where some degree of explicit control is apparent. We will see below that some schools of music analysis have developed an outlook towards music that is helpful in doing prosodic analysis. Some linguists have already used standard musical notation to describe linguistic events.

It has often been said that the crucial differentiation between music and the random sounds of nature lies in the fact that the music has been given structure by some human creator. While ideas of what is the appropriate analysis for this structure and how much structure must be imposed in order to have music have been disputed in the musical world, the necessity of some degree of structure expressible in some (musical) analytical terms is well founded. Bernstein, in his Harvard Lectures, attempted to give his view of nature of structure in music. He illustrated his lectures with musical examples from most periods, Classical, Romantic, Atonal, Twelve-tone, etc., especially comparing Schoenberg's approach to structure with Stravinsky's. Bernstein's own basic attitude towards the subject of structure in music is expressed in the following quote.

> A piece of music is a constant metamorphosis of given material, involving such transformational operations as inversion, augmentation, retrograde, diminution, modulation, the opposition of consonance and dissonance, the various forms of imitation (such as canon and fugue), the varieties of rhythm and meter, harmonic progressions, coloristic and dynamic changes, plus the infinite interrelations of all these with one another. These are the meanings of music. (p. 153)

Bernstein explicitly states in his introductory lecture that he considers music and speech to be linked to each other(p. 7 and later). We find that his conception of the interrelationship is at too far a remove from actual questions of analysis (either musical or linguistic) to be of much benefit to this account. His comparison of music and speech is, put very simply, that since speech links meaning ("deep structure") and sound ("surface structure") through purely linguistic structural analysis, we can do the same for music. We can capture the "meaning of music" by referring to the levels of structure. He also links poetry (as a superior level of structure) to aesthetic musical constructions. His diagram (p.85) comparing language and music is reproduced below, Figure 11.

LANGUAGE	MUSIC
D. Super-surface structure (Poetry)	D. Surface structure (Music)
C. Surface structure	C. Deep structure
B. Underlying strings (deep structure)	B. Underlying strings
A. Chosen elements	A. Chosen elements

Figure 11. Bernstein's schema relating language and music.

We are not convinced that the analogy which Bernstein proposes between language and music is the most appropriate one. More important for our purposes is Bernstein's perception of musical structure in its own terms. We will return to this shortly.

A major change in music theory came about when Schenker proposed that music should be analyzed in terms of the unfolding of basic tonal patterns that express the tonality. The rationale for this type of musical analysis has been given by Salzer (1962):

> The sole purpose of this kind of explanation ... is to show tonal coherence in gradual stages -- which is necessary, if we are to explain in a systematic way what we hear. (p. 23)

The goal of musical analysis, as expressed above, corresponds to a linguist's goal in explaining prosodic phenomena. Essentially Schenker's idea was that each piece of music, no matter how extended in time, was the elaboration by repetition of one of the few simple patterns which established a tonal center.

> Even where an artist succeeds in associating his ideas or images in a still more complicated way, still the principle of repetition can be recognized as underlying even the most daring feats. (Schenker: p. 11)

A piece of tonal music could be said to be in a given key because its melodies and harmonies elaborated a pattern which focused on that key. The elaborations were constrained, in well constructed music, to be additional instances of the basic pattern which underlay the piece. For this type of analysis, each piece of music is "an organic whole, in which every chord and motive has its meaning in the construction of the entire organism." (Salzer:p. 38)

In the same sense, the real connection between music and speech as instances of structured sound is reflected in Bernstein's analyses of various musical works. The elaborations of a segment of Mozart's G Minor Symphony (No. 40, K. 550) are shown to be hierarchically arranged repetitions of a two measure phrase (Figure 12, Bernstein (1976), p. 103).

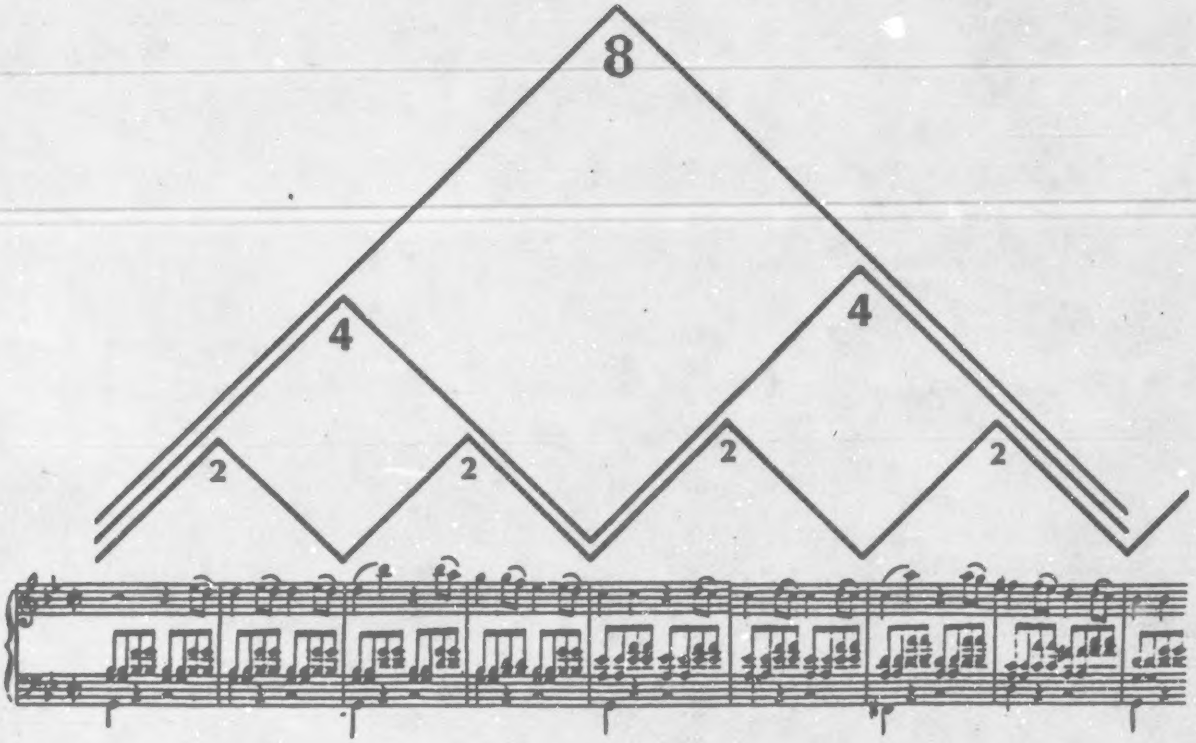


Figure 12. A Mozart symphony as elaboration of one phrase.

Another example of how Bernstein's view of musical structure supports our claims about tonal coherence comes in his discussion of Beethoven's Sixth Symphony (in F Major, Op. 68). There he performs an analysis of the piece designed to illustrate its tonal coherence. We offer part of his introduction, by way of summary.

> These first four bars are the material out of which the whole first movement is going to grow, an it's not just the main material but the only

> material. Every bar and phrase to come — and I mean everyone, without exception -- will be some kind of transformation, some metaphorical rendering, of the elements present in these four little bars. (pp. 157-9)

The relationship of this type of analysis of musical structures to questions about speech is straightforward. In proposing the existence of a limited number of tone groups, we are not denying that they are composed of individual tones. In the same way, Schenkerian analysis of music does not deny the existence of individual notes. Our proposal instead is that the unit of analysis should be these tone groups, and that all humanly possible sentences have to be elaborations of our small set of units.

> The idea is not new that mankind's creations, every complex thought or spiritual and artistic product is the elaboration, expansion, individual shaping or, in short, a series of prolongations of a basic underlying idea or structure. (Salzer:p. 28)

Thus our conception of tone groups becomes a part of the claim that speaking is an art[18].

II.4.2 Other possible arguments

We cannot, of course, totally rely on the analogy to musical analysis as a defense of tone groups. Without presenting a full motivation for the following, we would like to suggest some predictions which a tone group analysis might make.

---

[18] We are proud of our humanitarian bias in making such a claim.

We have already mentioned that our tone group proposal is a reasonable way to limit the number of possible basic tunes in the human repertoire. For musical analyses, Schenker's system allows only three basic tunes with a few additional modes of elaboration. These tunes are at most four notes long. We might claim that for language, no greater length is necessary. Our "declarative fall" basic tune required only three tones. If we limit tunes to three tones in length, with three tone levels, there are only eighteen possible tunes.

The second important implication of tone groups comes from the notion of elaboration by repetition (or modified repetition). The tone group demands elaboration in any except the simplest sentences. We have seen the syntactic dependency of the elaboration of the basic tunes in our analyses above. It is conceivable that rhythmic considerations[19] can be tied in to the notion of tune elaboration we are utilizing here. Incidentally, we are not limited to sequential repetition as the only method of elaboration. There is also what might be called parallel repetition, where each tone of the original tune is used as the basis for another instance of the basic (or elaborated) tune. In Medieval chants, this was called melisma, but we saw it occur in our analyses of complex sentences in the relationship between constituents and within a constituent.

Bolinger (1965, p.52) has pointed out a case where the vowel of "no" is repeated to indicate a pointed denial. The tone pattern he describes is "three successive A accents" which can be represented in our notation by:

---

[19] cf. Liberman (1975) for more discussion of rhythm in English prosody.

```
                        no - o - o
                        MHL  HL  HL
     or
                        MHL MHL MHL
```

Our analysis of this example is straight-forward. The basic tone group involved is what we have called the "declarative fall". The elaboration of that tone group is by repetition. If we accept the second representation of the A accent then we need not worry about the optionality of the mid tone. Otherwise, we consider the initial mid tone to be merely a "leading tone" to the rest of the contour, i.e., the mid tone is not part of the basic tone group, but only an elaboration of it. Thus, the elaboration could be represented as follows (G represents the tone group.):

```
                          no - o - o
                        M - G   G   G
     or
                            G   G   G
```

Finally, even in cases which are best regarded as sequential repetition of the basic tune, our analysis accounts for the variation between instances of the tune. We have noticed that the mid tone had to be designated an optional element of the autosegmental melody. The notion of a tone group suggested here would make it reasonable to eliminate the mid tone in certain, definable, syntactic situations as a form of elaboration of the basic melody (downstep). If we say that the mid tone acts like a leading tone in these sentential elaborations as well as in the lexical ones, then it is reasonable that only certain elaborations of the tone group will permit a leading tone, e.g.[20]

[[John] [stayed home]].
G G
154 128 - 127

[[John] [is [staying home]]].
G M - G
133 115 125 123

II.5 Emphatic stress: a result of tonal elaboration

The experimental evidence for this section is presented in depth below, but it is worthwhile to develop the theoretical framework for those results at this point. The basic finding is that emphatic stress can be simulated by changing the pitch contour, lengthening the word in question, loudening it somewhat, and surrounding it with silences. We can neglect the loudness question; lengthening the word and silences are similar and can be attributed to creating a percept of isolation which serves to highlight the utterance. The rules governing silences and word length seem to be very regular and fairly simple. The perceptual effect of modifying the pitch contour is also to create a highlighted, isolated area. The difference lies in the fact that changing the pitch contour is not simply local to the segments in question. Whereas we can lengthen locally, i.e., only the word in question is lengthened, we must modify a major portion of the pitch contour, and not just the word in question, in order to achieve the percept of stress.

---

[20] These examples are reported on elsewhere in this dissertation in more depth.

What is particularly nice about the use of tone groups for explicating this type of stress contour is that the non-local nature of the stress contour is represented by a local, limited perturbation of the tone group -- before elaboration! Suppose we link prominence, as used in autosegmental terms, with the marking of stress. Then, our elaboration of the tone group will force the prominence to the stressed word, and then the rest of the elaboration will operate naturally[21].

As a footnote to this section, we can notice that the neutral (non-prominence) elaboration of a tone group such as "declarative fall" will put its prominent tone on the first major element in the sentence. This explains why subject position is a favored perceptual position. In fact, a language specific fact about English is that words and phrases in first position in the sentence receive (without any special mechanisms other than the ones that put them in that position) emphasis. Such rules as Topicalization, Passive, and Clefting are syntactic mechanisms which serve to put a word or phrase first in the clause. Once the phrase has been moved to that position, no special emphatic stress need be added to the term to have it perceived as emphasized. The tonal pattern assigned due to its being in subject position is sufficient to give a listener the impression of stress. This phenomenon of subject stress is not true for languages which assign tone differently. French, for example, uses contours where the prominence is at the end of the word or phrase. In French we have examples where the subject is hard to stress unless it is embedded into a phrase final position. Compare:

---

[21] This is not exactly the implementation we have developed for stress, but it is ideologically homeomorphic to it.

/
? Jean est venu.

vs.

/
C'est Jean qui est venu.

II.6 Another derivative result

An independent result which can be derived from the tone group approach to intonation is a reformulation of the description of contraction phenomena in English. Contraction occurs, for example, in cases such as the following.

He is here. ==> He's here.
John would have gone home. ==> John'd have gone home.
or ==> John would've gone home.
or ? ==> John'd've gone home.

In general, reduced forms appear before the words they modify or belong to (Selkirk (1972), p. 40). Selkirk (1972) has explicated the nature of stress reduction and contraction and shown it to be related to the syntactic structure of the utterance. The strength of constituent boundaries determines the possibility of the stress reduction necessary for contraction to take place. Selkirk's system using word and constituent boundaries (taking into account the derivation of the surface constituent structure from a deep structure) captures the generalization that correctly describes the phenomena but it is stated in a formal system that leaves unexplained why the constituent structure and transformational history should have any bearing.

Our tone group approach also relies on syntactic structure, though not on the transformational history of the sentence, and can describe the same contraction phenomena even though these phenomena were not originally data for this approach. We explain contractions in terms of the tone contour assigned to an utterance. In our approach the tone contour, and therefore the possibility of contraction, is assigned based on the surface syntactic structure.

In applying our tone group approach to this problem, we are not finding new generalizations; we are only re-expressing the ones that Selkirk found. Where we feel there is an improvement over Selkirk's system is that our expression of these generalizations is in terms that are somewhat more explanatory. Furthermore, while Selkirk's system requires the phonology to have access to the transformational output, in the form of traces of deleted elements, our system works directly from the surface syntactic structure. Also, by applying our methods in a domain for which they were not originally intended, we will show the general applicability of our approach to phonological problems related to intonation.

II.6.1 Stress reduction and contraction

Stress reduction in English is a phenomenon where minor lexical[22] monosyllabic items lose their stress and can optionally be coalesced with a preceding word. A minor lexical item is a word from one of the classes: auxiliary verbs, prepositions, pronouns, determiners, complementizers. The major lexical classes are: verbs, nouns, adjectives,adverbs[23].

The basic fact to be explained about stress reduction is that reduction doesn't occur before a deletion site. A deletion site, in transformational terms, is the location of a word or a sequence of words that was generated in the base phrase marker but which has come to be deleted in the course of the derivation of the surface string. In trace theory, a deletion site would be marked by a trace of the "deleted" element. The trace may actually have been inserted as part of the deep structure and interpreted as coreferent to some other string or may mark the deep structure position (or "trace") of a deleted, moved or copied constituent. The sentences in Figure 13 (from Selkirk (1972), p. 55) are examples of the blocking of stress reduction when the minor lexical item is immediately before a deletion

---

[22] Selkirk uses 'non-lexical' where we have 'minor lexical'.

[23] In the synthesis system which is discussed below, minor lexical items receive a stress level of either five (for prepositions and complementizers) or six (pronouns, auxiliary verbs and determiners) on an eight level scale. Verbs and nouns receive a stress level of three and adjectives receive a four. These eight levels are parallel to the four levels used by the Sound Pattern of English; one stress is the most stress available and eight stress is the least (although not necessarily indicating a reduced vowel).

1. I asked where she is ___ now.
*I asked where she's ___ now.

2. You can dance in NY and you can ___ in Paris, but you'll never dance in LA.

Figure 13. Examples of non-reduction before a deletion site.

site. The deletion site is marked in the figure with an underscore (___). The underlined words are minor lexical items that would normally undergo stress reduction. In sentence one, the deletion site is due to the movement of the word 'where' from a deep structure looking like:

$[_S$ I asked $[_{NP}$ $[_{S'}$ COMP $[_S$ she is where] now]]]

to a surface structure in which 'where' is moved to the COMP leaving behind a trace of its deep structure location.

In sentence two, the word 'will' in a position syntactically similar to the underlined 'can' is reduced and undergoes contraction. The first occurrence of 'can' in the same sentence has a weaker stress than the underlined occurrence. The difference between the stressings of these syntactically similar uses of the minor lexical items is that the underlined word occurs just before the location of a deleted verb, 'dance'. A normal transformational history of the sentence would show three clauses each containing the main verb 'dance'. The verb is deleted in the middle clause[24].

---

[24] From our point of view, the difference between describing the transformational history in terms of a syntactic deletion leaving a trace, and surface interpretation of a deep structure trace is irrelevant to the present discussion.

Selkirk proposed a 'Monosyllable Rule' to account for the cases of destressing leading to contraction. We reproduce her rule (Selkirk (1972) p. 54) as Figure 14. The # is a constituent boundary. It's function in the Monosyllable rule is to force the minor lexical monosyllabic word (represented by $[C_0 \_\_\_ C_0]$ above) to become destressed when it is part of a constituent with a stressed major lexical word $(X \underset{[+stress]}{V} Y (\#)]\ Z)$ Selkirk's system prevents destressing before a deletion site by using major category boundaries that are distinct from minor category boundaries and which are not disturbed by movement and deletion rules (generating their own traces as it were). Thus Selkirk's rule depends on having a transformational history marked on the surface structure in order for her rule to achieve correct results. In the sentences one and two above, the deletion (or movement) leaves no stressed word inside of the same constituent, while preserving the constituent as a unit. According to the Monosyllable rule, no destressing of the monosyllable can occur, which is the correct prediction.

$$V \rightarrow [-stress]\ /\ [\ \#\ W\ [C_0 \_\_\_ C_0]\ [(\#)\ X \underset{[+stress]}{V} Y\ (\#)]\ Z\ \#]$$

Condition: X is not T##U .

Figure 14. Selkirk's monosyllable rule.

The syntactic bracketing which is used for tone assignment in our approach is dependent on the surface structure of the phrase. This bracketing is cognizant of the deletion sites in these cases since the reducible elements occur as the first string in their constituents, and if the head of the constituent is deleted the rest of the phrase will be empty. Thus when a tone group is assigned to that level of phrasal analysis, the optional mid tone can't be assigned since the condition on assignment is that the element not occur alone in its phrase[25]. Given this analysis of mid tone assignment, we can make stress reduction dependent on having an unstressed (really low stressed) mid tone absorbed into a "major" HL contour. Where there is a deletion site, the surface structure won't allow the preceding monosyllable to receive a mid tone and thus there will be no 'absorption of contours (single words are not bracketed in the examples below). See Figure 15 and Figure 16.

I asked where she is ___ now.

[I [asked [where [she is] now]]].

```
      (M) H  H'       H''L
          (M)H H'L
```

Figure 15. A low stress NON-mid tone retains stress.

A slightly more complex example is given below in Figure 17. In order to maintain our claim that only the surface structure

---

[25] The restriction on mid tone assignment was motivated independently from any facts about stress reduction.

I asked where she is staying ___ now.

[I [asked [where [she [is staying]] now]]].

```
        (M) H  H'                  H''L
               (M)H   H'         L
                    M    H    L
```

Figure 16. A low stress mid tone reduces stress.

a) John knows where the concert is ___ at four o'clock.

[where [[the concert] is] [at four o'clock]].

===>

*John knows where the concert's at four o'clock.

b) John knows that the concert is at four o'clock.

[that [[the concert] [is [at four o'clock]]]].

===>

John knows that the concert's at four o'clock.

Figure 17. A more complicated example of stress reduction.

of the utterance and not the transformational history is important, we have to explain why the phrase 'at four o'clock' is bracketed differently in the two different sentences above. If we can explain that satisfactorily, then the impossibility of contraction in (a) will be explained on the same basis as in the previous example, above.

We could explain the different bracketings of 'at four o'clock' as being reflections of the difference between locative and time adverbial uses of the phrase. The locative phrase is normally bracketed with the verb phrase in syntactic structure rather than with the sentence as a whole. The time adverbial, on the other hand, is generally bracketed with the entire sentence. There are exceptions to this structural generalization. For example in the sentences below, the same time adverbial is forced into different parts of the constituent structure simply in order to maintain the general tree configuration, yet there is no need to claim that the time adverbial changes its syntactic function in each of the sentences.

> The concert is in the main cafeteria at four o'clock.
> The concert is at four o'clock in the main cafeteria.
> The concert at four o'clock is in the main cafeteria.
> At four o'clock, the concert is in the main cafeteria.

The different positions open to adverbials is a well known and hard to account for fact about English. It is well known that the by-phrase of passives acts like an adverbial (it has been associated with manner adverbials) and has a relatively free positioning somewhere after the main verb though its normal position is at the end of the sentence, i.e., it is realy adjoined to the sentence and not to the verb phrase.

> John was arrested on the corner by the police.
> John was arrested by the police on the corner.

There are still some questions to be answered in substantiating the claim that the phrase 'at four o'clock' is sometimes an adverbial and sometimes a locative. How do we know which instance of the phrase we have in any given sentence? The answer to that comes in two parts. First, notice that in the (a) sentence there is a locative phrase which has been preposed (where), while in the (b) sentence there is no such locative. This means that in the (b) sentence there is a locative position available to the structure (unfilled by lexical material or traces) which is not available in the (a) sentence since there is already a locative, albeit moved away from its normative position. Thus it is structurally possible to call the at-phrase a locative in (b) but not in (a). The semantics of the (b) sentence indicates that the at-phrase is acting as a locative in that sentence, in the sense that it locates (temporally, rather than spatially) the event in question. The proposition expressed by the (b) sentence is that John knows about a concert-at-a-given-time rather than that he knows that the concert is and that it is at a given time. The at-phrase modifies the noun phrase, the concert in (b) rather than modifying the proposition that there is a concert. Note that in (a) the at-phrase is modifying the proposition that a concert is somewhere. The reading with the at-phrase modifying the noun phrase would yield a proposition that a concert-at-a-given-time is somewhere.

Notice that in the following sentence, if it is grammatical, there is the possibility of stress reduction on is.

? John knows that in the cafeteria the concert is at four o'clock.

? John knows that in the cafeteria the concert's at four o'clock.

In a system which maintained a transformational history, it would be claimed that the in-phrase was preposed from a position after the at-phrase, thus allowing contraction of the is exactly as in the sentence:

John knows that the concert is at four o'clock in the cafeteria.
John knows that the concert's at four o'clock in the cafeteria.

In a system that only looked at the surface structure of the utterance we would have to rely (somewhat lamely) on the semantic explanation of the function of the at-phrase in these sentences in order for the bracketing to come out correctly for the tone assignments.

There is a class of sentences for which the reasonable surface bracketing gives the wrong predictions. These sentences are cases of unbounded movement or deletion of a constituent, e.g., "Tough"-Movement or Comparative Deletion. Selkirk's theory achieves the correct predictions through use of her "Traces" of the moved/deleted element. For example we have the following contrast.

(a) Eager though I am to win, I would never cheat.
vs.
(b) * Eager though I'm to win, I would never cheat.
vs.
(c) Though I am eager to win, I would never cheat.
vs.
(d) Though I'm eager to win, I would never cheat.

In transformational terms, the word 'eager' has moved from being immediately after the word 'am' to the beginning of the sentence. We see that the phrase [am [eager [to win]]] is a constituent and it is hard to argue that [am ___ [to win]] is also not a constituent. It is certainly a constituent in the simple sentence 'I am to win.' It would certainly be a case of arranging the syntactic structure to yield the correct phonological results if we were to assign 'am' and 'to win' to separate constituents.

It seems that in order to handle such cases we need to resort to some mechanism which will enable us to identify 'missing' constituents. We could do this by incorporating a 'trace' marker into surface structure, or by encoding a transformational history, or by utilizing an additional level of syntactic, semantic, and possibly other, information to determine the surface syntactic structure[26].

We might note that our very simple account gives the correct predictions for contraction phenomena in a wide variety of cases. All of these cases are "simple" when a comparison of corresponding transformational rules is used. By that sort of standard, only movement or deletion rules which involve a 'crucial variable' (Ross 1967) are not "simple". Still, while our account has not discredited Selkirk's work in this area, we have provided a framework that gives some additional motivation for why her rules work as they do and which provides an indication as to how necessary the devices (especially traces) she introduces are.

---

[26] Langendoen (1975) argues for rebracketing rules which could probably be used to achieve the same result we want here. These rebracketing rules, themselves, would need to be motivated by syntactic, semantic or other criteria.

Chapter III

Experimental results

Let us assume that our vocabulary of constructs and processes is suitable for describing English intonation. We next need to discover the specific forms that these constructs take on in English. What tone groups are utilized in English? How are these tone groups elaborated? How does English use the elaborated autosegmental specification of an intonation contour to generate an actual phonetic pitch contour? To answer these questions we will examine the simplest examples first and then proceed to more complex cases. Along the way we will enunciate and then modify a series of generalizations which will encompass some answers we are looking for.

III.1 Simple sentences

In this section we will present a discussion of declarative prosodic contours, based on evidence from the simplest sentences in English. Examining simple declarative sentences, we find that there is a general phenomenon we will call "downstep". In the simplest two-word sentences and proper names, the first word is consistently higher in pitch level than the second; i.e., the first word consistently downsteps to the second. Downstep can be viewed as the instantiation of the 'High-Low' tone group which we will give the name "declarative fall". Thus we have the following pattern[1]:

---

[1] The pairs of pitches, corresponding to two word sentences or

WORD	PEAK FREQUENCY		WORD	PEAK FREQUENCY	
	(a)	(b)		(a)	(b)
people	256	154	rabbits	238	123
cursed	252	123	multiply	220	116
Pamela	154	173	Pamela's	232	130
Jackson	147	169	dreaming	215	112
Abernathy's	122				
tired	114				

Figure 18. Two word sentences and proper names.

In slightly longer sentences[2] and proper names there are the following distributions of peak frequencies (In Figure 19, The speakers of "people cursed Nixon"(a), and "Richard Milhous Nixon"(c,d) were female; the speaker of "please type the answer" (both readings had the same reader) had a higher (male) voice than the male speakers of the other sentences.):

---

proper names, were recorded by a range of speakers, male and female. The (a) and (b) columns in each case were not recorded by the same speaker. Some of the data presented was furnished by Will Leben.

[2] We are treating imperatives as if they have the same contour as declaratives.

WORD	PEAK FREQUENCY	
	(a)	(b)
people	244	133
cursed	229	128
Nixon	234 (226)	110
please	200	204
type	172	161
the	154	161
answer	143	133
green	160	158
dreams	144	126
sleep	125	118
peacefully	96	108

WORD	PEAK FREQUENCY			
	(a)	(b)	(c)	(d)
Richard	175	125	200	200
Milhous	150	120	160	177
Nixon	140	110	150	150
John	154			
typed	137			
the	115			
answer	120			

Figure 19. Three and four word sentences and proper names.

Notice that the first downstep for the sentences starting at 200 hz was about 30 hz while those for the sentences starting at lower frequencies were proportionately lower.

We conclude, provisionally, that we can superimpose a phonological (peak to peak) pitch contour consisting of a fixed downstep between lexical items. We would represent this downstep as the tone group: H L, which expands to an autosegmental description as: $\underset{0}{H}$ L. This downstep can be imposed, in practice, by multiplicatively adjusting all of the pitches in each word contour so that the maximum pitch in each one is at the assigned level. If we impose no other prosodic manipulations to such sentences, each synthesized word would sound as though it were final in the sentence. We will discuss in a later section (Section III.5) the various methods we have employed to achieve a realistic smoothing of the contour. Given these smoothing mechanisms and using the downstep elaboration, we can generate fairly natural sounding pitch contours for sentences from the concatenation of lexically recorded words.

III.1.1 A non tone group view

It has been proposed in the literature to generate a polynomial curve (up to 4th degree seems necessary) that, it is claimed, simulates the total melody of these simple sentences[3]. On this analysis the melodic curve is generated and then superimposed on the entire sentence, or perhaps used as a baseline for the frequency curves of the lexical items. It is reasonable to expect that there might be dependencies between the syntactic relationships of the words in a sentence and the coefficients of the prosodic curve. This would imply that sentences different only in the lexical items used, or their syntactic relationships, could have mathematically unrelated curves. It would be necessary, on this view, to have a substantial catalog of simple sentences arranged according to detailed syntactic pattern (and perhaps other factors: word stress, number of syllables, vowel qualities, etc.). The question of extending either view to longer, more complex sentences will be discussed shortly, but the most straightforward downstepping approach, if it worked, is much simpler than any n-th order polynomial curve.

---

[3] cf. Olive and Nakatani (1974), and Olive (1975).

A perhaps stronger criticism of the polynomial approach is that it is linguistically uninformative. It is conceivable that both accounts would be descriptively equivalent; they would make equally good predictions. We would still prefer the theory that is related to more general linguistic principles. Our account of English prosody was linked, above, to the analysis of tone languages. The challenge of linguistic relevance has to be met by the polynomial account. Simply to be able to describe intonation with polynomials (or other devices) is insufficient.

III.2 More complicated sentences

Let us proceed to the analysis of more complicated sentences. We will persist in our view that the analysis should be carried out in terms of (1) tone levels: abstracted from the continuous data as the maximum pitch of the words in the sentence, and (2) downstep: the relationship between the constituents of the "normal" declarative sentence.

III.2.1 Some data

We list below, in Figure 20, the pitch peaks for several complex sentences. The graphs of fundamental frequency and loudness for these sentences from which the data was abstracted are given in Appendix I. All of these examples were recorded by a single male speaker of college age at a single recording session which included many other sentences in a somewhat randomized list. The speaker is the same person who recorded the individual words in our lexicon.

WORD	PEAK FREQUENCY	WORD	PEAK FREQUENCY		
1.		2.	(a)	(b)	
read	213	choose	200	208	
the	192	the	192	164	
text	167	line	154	152	
if	164	that	123	135	
you	175	implies	159	143	
can	156	line	141	137	
		ten	145	141	
3.		4.	(a)	(b)	(c)
the	115	rule	208	204	208
rule	200	aa	172	154	172
can	185	requires	135	133	147
be	118	two	175	182	181
found	168	lines	151	172	179
on	133	one	175	172	167
page	159	of	172	167	164
seven	143	which	179	172	170
		must	149	143	149
		be	145	147	159
		a	135	143	152
		conditional	151	123	161

Figure 20. Complex sentences.

On a surface scan of this data, it would seem that our hypothesis concerning downstep has been invalidated. In each of the sentences there are some transitions that are clearly not downsteps. We can no longer map our tone group, H L, directly onto the autosegmental tone sequence H L and then to the concrete tones H_0 H-1 H-2 ... L. The surface contour is no longer a straightforward downstepping of tones. Yet we maintain that the regularity of downstep is, in fact, present in this data. We have to introduce several different modes of elaboration in order to crystallize the regularity of downstep in these examples.

III.2.2 Observation I

We observe that sentence 1 deviates from the downstep between "if" and "can". But if we bracketed the sentence into its two clauses we would have

[[read the text] [if you can]].

Looking only at the maximum frequency in each clause we find that there is a downstep between them from 213 hz to 175 hz. Furthermore, within the first clause there is the normal downstep relationship between the words. In the second clause we see the downstep relationship except for the word "if". We will later claim that words like "if" normally upstep (in a way to be made precise later) when they occur at the beginning of a clause. Given this secondary claim, we can proceed to analyze sentence 1 on the basis of the following generalization[4].

III.2.2.1 Generalization A

> Prosodic processes (in particular, tone group elaborations) are sensitive to the phrase structure (at least to phrase boundaries). In the elaboration of the "Declarative fall" tone group (tentatively: H L) there is a downstep relationship between clauses and another, distinct

---

[4] Selkirk (1972) makes a similar generalization based on evidence from the analysis of word-word contractions. See above Section II.6.

> downstep within a clause. The downsteps within a clause are linked to the downstep between clauses so that the highest pitch in each clause′s downstep is the appropriate pitch in the downstep sequence for the clauses.

Generalization A yields an analysis of sentence 1 which agrees with our intuitions that separate clauses should have separate, though linked, prosodic contours.

III.2.3 Observation II

If we look at sentence 2 above we find that there is a regular downstep from "choose" through "line" (and perhaps extending to "that"), and another downstep from "implies" (H′ is the lowered high tone initiating this downstepping sequence) through "ten" with the word "line" being a small abberation (designated as "H-?") from the straightforward downstep. An autosegmental description of this sentence might be[5]:

Choose the line that implies line ten.

	H	H-1	H-2	H-3	H′	H′-?	H′-1	— L
(a)	200	192	154	123	159	141	145	(85)
(b)	208	164	152	135	143	137	141	(87)

This sort of description might lead us to believe that the intonation contour for a sentence is generated from a series of downsteps, i.e, repetitions of the H L contour. The extent of each series of steps could be attributed to a construct like "breath group". In the above example, we would postulate two breath groups, the first from "choose" to "that" and the second from "implies" to "ten". The final low tone is due to sentence final intonation.

---

[5] (a) and (b) are the pitches given above in Figure 20.

Choose the line that implies line ten.

|H L| |H' L|

In fact this view could lead us to at least one incorrect prediction. If the downstep is viewed as H L then there should be a low tone at the end of the first breath group as well as sentence finally. We could postulate a "Low tone absorption before High" rule to account for the difference or we could postulate another structure, such as is given below.

Choose the line that implies line ten.

| H H H H| | H H H-L|

The prosodic rules used for this version of the analysis would be:

1) Assign high tones to all the words in a sentence and a sentence-final low tone to the last word.

2) Make the downstep a context sensitive rule, dependent on surrounding tones and breath phrasing.

More ad hoc additions to the analysis could be made as necessary, but it should be relatively clear that the basic premise of this approach depends on being able to define a breath group reasonably. Any such definition would seem to be problematic in this case.

If we were to give a syntactic parsing to this sentence, we could bracket it either as

```
        [ choose [ [ the line] [that implies [line ten] ] ] ].
or
        [ choose [ the [line [that implies [line ten] ] ] ] ].
```

We have either a nominal ("the line") with a relative clause, or, an embedded relative clause related to the noun "line". The phrase structure rules for the two cases would be:

```
        NP --> NOM + S;          NOM --> (DET) (ADJ) N ...
or
        NP --> (DET) (ADJ) NOM; NOM --> N + S ...
```

We can successfully analyze the relative clause using generalization A above and repeating the claim that certain words upstep at the beginning of a phrase. The pattern of downstepping becomes clearer if we list the frequencies in terms of the above bracketings.

```
        [ choose [ [ the line] [that implies [line ten] ] ] ].
        [  200   [ [ 192  154] [ 123  159    [ 141 145] ] ] ].
        [  200   [      192    [ 123  159         145   ] ] ].
        [  200   [      192              159              ] ].
or
        [ choose [ the [line [that implies [line ten] ] ] ] ].
        [ 200    [ 192 [ 154 [123   159    [141  145] ] ] ] ].
        [ 200    [ 192 [ 154 [    159          145    ] ] ] ].
        [ 200    [ 192                 159                ] ].
```

Once again we can see that there is a normative downstep through the constituents of the sentence and within each constituent. If, instead of considering the full constituent structure(s), we simply looked at the word "line" and the relative clause we would find that there is a small upstep between their tone levels (154 and 159).

This finding could lead to an assumption of a special relationship between the tone levels of a noun phrase head and its complement; some special small upstep. Yet, in the other recording of this sentence there is a downstep between for these elements (152 and 143). On the NOM + S analysis any relationship between the head and complement of a noun phrase is totally spurious. If we write downstep as a function,

D("a tone level", "level of embedding"),

we can see that the relationship between head and complement is accidental. Assume that the level for the whole phrase is H. Then the tone level for the NOM is H, and the tone level for the head noun is D(H,2). The tone level for the complement is D(H,1). Unless we can fix some relationships between the downsteps at different levels of embedding, there is no necessary connection between the two downsteps. We could also apply the second syntactic analysis to this issue as well, but since the argument would largely parallel the one already given here, we omit it.

Notice, further, that on either analysis presented, it is only accidental that "that" is downstepped from "line" since the frequency for "that" is derived from the frequency of the relative clause and not from that of the noun. Many other cases of seeming downstep are only offshoots of the syntactic structure of the sentence, just as many apparent exceptions to the rule of downstep are.

We can summarize what we have done in the analysis of this sentence as Generalization B.

III.2.3.1 Generalization B

> The phrase structure on which the intonation assignment is dependent is the surface syntactic structure. Furthermore, the depth of embedding of a constituent in the structure will determine the size of the step being actualized in elaborating the contour to that constituent.

III.2.4 Observation III

Sentence 3 can be almost completely analyzed on the basis of the "declarative fall" tone group and generalizations A and B. If we look at a syntactic parsing of the sentence we get:

[[the rule] [can [be found] [on page seven]]].

We have given the sentence an analysis as NP+VP, where VP is expanded as Aux+V′+PP (V′ represents a be-verb+adjective or be-verb+participle node). As before we will rewrite the frequencies in bracketed phrases.

```
[ [the rule] [can [be found] [on page seven] ] ].
[ [115 200 ] [185 [118 168 ] [133 159  143 ] ] ].
[ [  200   ] [185     168          159       ] ].
[    200              185                      ].
```

We proceed to state generalization C which expresses our interpretation of the above case.

III.2.4.1 Generalization C

> The "declarative fall" tone group

elaboration is sensitive to syntactic structure in such a way that both phrasal nodes and major lexical nodes immediately dominated by a common node downstep with respect to each other, but minor lexical nodes need not downstep when in phrase initial position.

We can apply this generalization to sentence 3 as follows: the bottom line in the analysis above represents the frequencies of the phrase nodes for NP and VP; the line above that shows the constituents which are immediately dominated by VP; within each of the constituents V' and PP we have the upstep from the first element ("be" and "on", respectively) to the rest of the constituent. This is similar to what we noticed in sentences 1 and 2; we will concretize our understanding of the upstep later. Within the PP we have a downstep between "page" and "seven".

Generalization C lends itself to two interpretations. One view might be that minor lexical items do not participate directly in what we have labeled tone group elaboration. They are assigned their pitch levels, on this account, through lower level phonetic or phonological processes or rules. The tone elaboration for this sentence actually looks like the following:

```
[ [the rule] [can [be found] [on page seven] ] ].
        H                   H-1                 L
        H      H'     H'-1         H'-2         L
        H      H'      H'-1       H''  H''-1    L
[ [115 200 ] [185 [118 168 ] [133 159  143 ] ] ].
```

The words "the", "be" and "on" do not receive any tone assignment at this stage of analysis because they are minor lexical items. A separate context sensitive rule will assign the appropriate tones to these words. We have used the apostrophe (and negative integers) to designate a downstepped tone in this diagram for convenience; we could have written the tones out in terms of the downstepping function, D(T,E), introduced above.

An alternative interpretation of generalization C for this sentence is that the sequence of tones representing the tone group, "declarative fall", is actually "optional mid tone, followed by a high tone and then a low tone", written as (M) H L. Let us give the alternative analysis below.

```
[ [the rule]   [can    [be found]   [on page seven] ] ].
        H                         H-1                  L
   M    H         H'      H'-1           H'-2          L
   M    H   L     H'    M' H'-1 L'  M'' H''  H''-1 L''L
[ [115 200 ]   [185    [118 168 ]   [133 159  143 ] ] ].
```

According to this analysis, we would expect that the different mid tones assigned, M, M', M'', would each bear a different step size with respect to the principal high tone in that phrase. Just as the step size for the downstep between constituents is different at different levels of embedding, we expect the upstep size to be similarly dependent. We could thus define the mid tone as a function of the high tone and level of embedding (step size), M("high Tone", "level of Embedding"). We would then have:

M(200,1)=115, M(168,2)=118, M(159,2)=133.

As the pitch decreases and the level of embedding increases, the step size, from mid to high, decreases. Thus the analysis which includes an optional mid tone in the tone group predicts the placement of the low tones, which would need to be inserted by a separate rule on the other analysis. We will henceforth consider the "declarative fall" tone group to be (M) H L.

III.3 Some complicated simple sentences

In this section we will concentrate on the "mid to high" part of the "declarative fall" tone group. We present evidence concerning phrase-initial upstepping of certain types of words which we regard as exemplifying this part of the tone group. While we will be concentrating our attention on the auxiliary and verb phrases, it is important to recall that under our hypotheses, each phrase, regardless of its label, exhibits parallel structure in regard to the elaboration and mapping of tones from a tone group, and, in particular, to the downstep and upstep relationships.

III.3.1 Data

WORD	PEAK FREQUENCY	WORD	PEAK FREQUENCY	WORD	PEAK FREQUENCY
1. (a)		(b)		(c)	
John	133	John	135	John	141
is	115	may	128	may	139
staying	125	be	110	have	135
home	123	staying	116	been	114
		home	114	staying	120
				home	119
2. (a)		(b)		(c)	
John	154	John	137	John	167
stayed	128	may	128	may	147
home	127	stay	115	not	128
		home	103	stay	132
				home	110
3. (a)		(b)		(c)	
John	149	John	152	John	149
may	133	may	132	could	125
not	130	not	127	not	125
have	135	have	135	have	130
been	114	stayed	115	stayed	115
staying	123	home	114	home	108
home	119				

Figure 21. Complicated simple sentences.

These sentences were all recorded by one male speaker (who was not the speaker of the sentences in Figure 20).

III.3.2 A first principle

The data in group 1 above shows the upstep occurring on the morpheme "be" (in its different realizations) in each of the sentences. Except for that morpheme the downstep is regular for those sentences. Sentence 1a would normally be given a surface syntactic structure of Subject + Predicate, where a reasonable decomposition of the Predicate is Copula + VP, and VP is Verb + Adverb. We can give a similar structure to sentence 1b: Subject + AUX + Predicate, where the AUX node is filled by "may"[6]. Our surface structure for 1c is similar to 1b except that where 1b has only "may" in the auxiliary, 1c has both "may" and "have". For our current purposes the important aspect of the syntactic structure of the sentences in group 1 is that the morpheme "be" which is the first element of the Predicate upsteps to the rest of the phrase. We state this finding as Generalization D:

III.3.2.1 Generalization D

> In the elaboration of the declarative fall tone group, the first element of a major phrase node may receive the optional mid tone, i.e., upstep to the assigned pitch level for that node, if it is an unstressed morpheme and the major phrase node is sufficiently heavy[7].

We have stated the generalization in terms of the possibility of an upstep occurring, since in other data not shown here, we have found that a good deal of optionality is involved, perhaps dependent on semantic considerations. Furthermore, while we speak in terms of unstressed morphemes in the generalization, this is really a substitute for the term "minor lexical item" or "non-lexical item"[8]. The use of "unstressed" is slightly more appropriate than other possible terms, since we can understand stress as a semantically related term. For example, in the sentence discussed below, "Rule aa requires two lines; one of which must be a conditional", there is an upstep from "requires" to "two lines" in all three readings of the sentence. We can say that "requires" is acting as an unstressed morpheme, semantically equivalent to "be" or "must" in the reading that it has in this sentence. With the formulation of Generalization D as above, we can assign a mid tone to "requires" since it is unstressed. If we had formulated our generalization in terms of "minor lexical items" we would have to explain why some major lexical items could also receive a mid tone. We have seen above, in Section II.6, that the idea of minor lexical item has an important function in the grammar. Our tone group elaboration and mapping rules are more general in formulation than the rule for contraction.

---

[6] While other analyses might be proposed, we feel that the structure given above is reasonable enough to stand on syntactic grounds alone. For recent work on this topic, see Akmajian and Wasow (1975).

[7] While the term "heavy" employed above lacks precision, we note that this usage is similar to that employed in describing "Complex Noun Phrase Shift", the syntactic rule that states that a noun phrase can be moved rightward over some other constituent if it is sufficiently "heavy". Syntactically, heaviness is vaguely defined, as it can be predicated of long strings as well as of some short ones while it fails to be true of some long phrases. c.f. Ross (1967) Sec. 3.1.

III.3.3 Two possibilities

We can view the sentences of group 2 as having increasingly heavier auxiliary constituents. Sentence 2a has no auxiliary, and displays a paradigm downstep. Sentence 2b has the modal "may" in the auxiliary and can be analyzed as NP + AUX + VP with each constituent downstepping appropriately. On a similar parsing, sentence 2c has two elements in the auxiliary: "may" and "not". The upstep between "not" and "stay" is explained as resulting from the interaction of the downstep within the auxiliary and the downstep for the overall sentence (specifically from AUX to VP). We have already seen an instance of this "apparent" upstep in Section III.2.3.

---

[8] See above Section II.6 and the work of Selkirk, cited there.

Alternatively, a different parsing of 2c would adjoin the "not" to "stay" (the main verb). If we accept this second parsing, the upstep from "not" to "stay" is a real upstep and its explanation is the same as that given for the morpheme "be" in the group 1 sentences above. As we have presented them, we have no apparent motivation, either syntactic or phonological, for ascribing either surface structure to sentence 2c. Analysis of the sentences in group 3 will shed some light on this choice.

III.3.4 Clarification

In sentence 3a we have a more complicated situation. We can attribute a surface structure of NP + AUX + PRED to the sentence. Here the NP is "John", the AUX is "may not have", and the PRED is "been staying home." Within the PRED our analysis is straightforwardly the same as given in Section III.3.2 above. There is an upstep from "been" to "staying" followed by the regular downstep between "staying" and "home". According to our analysis of the auxiliary of sentence 1c above we should parse the AUX simply into three separate elements, i.e., "may + not + have". Under this parsing, we must explain not only why "not" upsteps to "have" but also why "may" upsteps to "have". Alternatively, we might rephrase the problem as why the mid tone for the phrase occurs in second position rather than phrase initially.

According to Generalization D the upstepping of "may" to the rest of the AUX (in particular: "have") is not totally unexpected. If we assign a pitch level of 135 hz to the entire auxiliary, then "may" is an upstep to that level. That upstepping reflects the heaviness of the AUX; it is easy to imagine that, with the addition of "not" to the AUX, the scales have shifted enough to allow the upstep. Now our problem is only to understand why "not" is at a lower level than "may" instead of a higher one as Generalization D predicts, i.e., why "not" does not get the high tone assigned to the AUX node. If we say that the internal structure of the AUX here is similar to the AUX of sentence 2c, we might assign it the following bracketing.

[[may not] [have]]

That is, we might suppose that the "not" is adjoined to the element preceding it in the AUX. In that case we need to change Generalization D from a statement only concerning single elements within a constituent to a rule that states that a complex constituent may upstep if it is phrase initial, relatively unstressed and occurs in a heavy enough phrase. Then the putative constituent [may not] is assigned an upstepping pitch, but within that constituent there is a normal downstep from "may" to "not". Changing our generalization from allowing upstep (elaboration of the optional mid tone) only on a single morpheme to allowing it on a constituent is a significant weakening of the constraints of tone group elaboration and we would like to avoid it if possible.

An alternative structure for the AUX would make it parallel to the second parse suggested for sentence 2c. We can adjoin the negative morpheme to the following AUX (or VP) element. Then the AUX would have the following structure.

[[may] [not have]]

With this analysis of the AUX we can apply Generalization D twice, once for the constituent "not have" and once again for the step between "may" and the rest of the auxiliary. With this analysis of the auxiliary we would not have to weaken the constraint on optional tone elaboration which is embodied in Generalization C. On strictly empirical grounds, there does not seem to be a way to choose between these two alternatives, but we will examine the rest of the data before giving up.

Sentence 3b simply duplicates the situation of sentence 3a, except that here the upstep of the morpheme "be" to the verb phrase is missing. In sentence 3c, however, we find that both "could" and "not" have the same pitch level. This fact is in apparent contradiction to the two other cases in group 3. Only with the second syntactic analysis can we give a satisfactory account of the data of 3c (if it is valid data): "not" is adjoined to the following element ("have" in 3a,b,c, and "stay" in 2c). In fact, equivalent pitch levels for the two morphemes is entirely reasonable given that two separate upsteps to the same (constituent) pitch level are taking place. In this vein, we remark that there is a possible experimental error, due to segmenting the spoken sentence, of several hertz for the word "may" in sentences 3a and 3b. This error is not present for "could" since the initial voiceless obstruent makes identification of the beginning of the word unmistakable.

We have thus found an intonation fact for which an analysis of the auxiliary that says that the negative morpheme is adjoined to the following verbal element is preferred. This analysis allows us to dispense with additional principles while extending Generalization D to straight forwardly cover a wider variety of cases. Finally, we note that a similar, and potentially clearer, situation arises with respect to multiple prepositions, e.g., "up in the air" where intonation facts similar to those presented here are found. In those cases, an analysis of the complex prepositional phrase as being PREP + PREP-PHRASE seems easily justified. That analysis corresponds to our second version of the structure of the auxiliary, the one we have adopted here.

III.3.5 Using the generalizations

As a sample of our prosodic methods at this point, we give below the syntactic analysis of sentence 4 (from Figure 20) together with a discussion of the three frequency analyses which were abstracted from three separate utterances (by the same speaker) of the sentence.

```
(A)
[[[rule aa][rq. [two lines]]]; [[one [of which]][must [be [a cnd'l]]]].
[[[208 172][135 [175  151 ]]]; [[175 [172 179 ]][149  [145[135 151]]]].
[[[  208  ][135  175       ]]; [[175    179    ][149  [145    151  ]]].
[[[  208  ][135  175       ]]; [[175    179    ][149      151      ]]].
[[   208     175            ]; [     179                  151      ]].
[           208             ;                 179                  ].

(B)
[[[rule aa][rq. [two lines]]]; [[one [of which]][must [be [a cnd'l]]]].
[[[204 154][133 [182  172 ]]]; [[172 [167 172 ]][143  [147[143 123]]]].
[[[  204  ][133  172       ]]; [[172    172    ][143  [147    143  ]]].
[[[  204  ][133  172       ]]; [[172    172    ][143      147      ]]].
[[   204     172            ]; [     172                  147      ]].
[           204             ;                 172                  ].

(C)
[[[rule aa][rq. [two lines]]]; [[one [of which]][must [be [a cnd'l]]]].
[[[208 172][147 [181  179 ]]]; [[167 [164 170 ]][149  [159[152 161]]]].
[[[  208  ][147  181       ]]; [[167    170    ][149  [159    161  ]]].
[[[  208  ][147  181       ]]; [[167    170    ][149      161      ]]].
[[   208     181            ]; [     170                  161      ]].
[           208             ;                 170                  ].
```

(rq. = requires; cnd'l = conditional.)

Sentence 4 is composed of two adjoined clauses and therefore, by generalization A, there is a downstep between them. We find fairly good agreement in the step between the clauses in the different samples. In (A) we have a 29 herz step; the step is 32 herz in (B); and, in (C) the step is 28 herz.

III.3.5.1 The first clause

Within the first clause, there are downsteps both from NP to VP (NP=[rule aa], VP=[requires two lines]) and within the NP. In the data we show, the steps between the two phrases in the first clause are, respectively, 33 herz, 32 herz and 23 herz. Notice that while the step between NP and VP in first clause of (C) is smaller than in the other examples, there is a consistency to that utterance in that it has a smaller step size in its second clause as well.

The NP, [rule aa], exhibits downsteps of 36, 50 and 36 herz. The object of the verb 'requires', [two lines], also shows downstep: 24, 10 and 2 herz respectively. In both of these cases the step size shows a fair amount of variation. We can say that differing semantic 'weights' are implicit in the different step sizes of these phrases. We will return to this explanation after examining the steps in the second clause.

III.3.5.2 The second clause

In the second clause there is a downstep between NP and VP (NP=[one of which] and VP=[must be a conditional]). These downsteps are 28 herz, 25 herz and 9 herz respectively. Notice the good agreement of the step sizes in the two clauses of each utterance: i.e., (A): 33, 28; (B): 32, 25; (C): 23, 9. Under our explanation of step size as being a function of both syntactic embedding and fundamental frequency (pitch) of the phrase, we expect that the second clause will have slightly smaller steps in corresponding syntactic positions. the lower pitch of the second clause will cause a slightly smaller step to give the perception of a musical interval of equivalent extent. We will see that sentence (C), while agreeing in broad outlines with the predictions which are verified for the other examples, does not fit closely with our predictions of step size in most instances.

In the subject NP, the partitive construction gives us an instance of a pair of upsteps. Most deeply embedded, we have the upstep from 'of' to 'which' (7, 5 and 6 herz). In sentences (A) and (C) we also have clear examples of an upstep from word to a phrase (steps of 4, 0 and 3 herz). We have not specifically constructed the syntactic structure in the partitive to predict the possibility for upstepping of the word 'one' to the phrase [of which]). The structure shown for the partitive phrase is the structurally motivated one, taken from standard syntactic analyses of phrase structure.

Interestingly, we also have the tonological evidence for a multiple upstepping syntactic structure in the verb phrase of this clause. Two of the sentences show an upstep from 'a' to 'conditional' ((A): 16, (C): 9). The same sentences show another upstep between the word 'be' and the phrase [a conditional] ((A): 6, (C): 2). Furthermore, sentence (C) also shows an upstep from 'must' to [be a conditional] (12 herz). Looking only at the data from sentence (C) we see a triple upstep leading to the head of the phrase. The upstep is not monotonic, of course. If we were regarding the contour simply as a single instance of M H L, there would be no good explanation for why the (C) contour actually has a fall from 'be' to 'a'. In our system, each instance of the mid tone is due to a separate elaboration of the tone group and therefore, there is a certain amount of freedom in the step size chosen for any particular elaboration. This allows the successive mid tones not to be monotonically rising.

We see that in the (B) version of the sentence there is no upstep between 'a' and 'conditional'. There is also no upstep between 'be' and the phrase [a conditional]. Additionally, there is no clear upstep between 'one' and [of which] (a step of 0 herz is ambivalent between upstep and downstep) in the same clause. What seems to be happening in this instance of the second clause is that the semantic 'weights' of the different words and phrases of the clause are different in (B) than they are in the other two instances of the sentence.

A further analysis of recorded text can be found in Appendix II, where several recordings of "The North Wind and the Sun" are analyzed and compared to the tone levels that are generated by the synthesis system.

III.3.6 Summary of generalizations

Let us recapitulate the empirical principles that we have employed in the above analyses.

1) Our most basic assumption was that sentential prosody is amenable to analysis as a sequence of level tones.

2) The second assumption we made was that, for declaratives, we could analyze these tone levels as being in a general downstepping relationship that is sensitive to surface syntax. An unelaborated form of the "declarative fall" tone group was displayed in two, three and four word sentences and proper nouns.

3) The association of the surface syntax with the elaboration of the "declarative fall" tone group (downstep relationship) was specified. We developed a general principle that any constituents that were on the same level of a syntactic tree downstepped from the first such constituent to the last. Within each such constituent the same process holds true; but there is no necessary connection between the tone levels within different constituents except as dictated by the downstep relationship governing these higher level constituents.

4) In the elaboration of the "declarative fall" tone group, there is an initial upstep (from mid tone to high tone) in a phrase if the first word in the phrase is of a certain syntactic/functional category.

Using these principles we were able to account for the intonational contour of a fairly complex sentence.

Could the other analysis we discussed above (Section III.1.1) have accounted for the same range of facts? We left the other analysis at the point where it seemed to require an n-th order curve, where n was the number of words in the sentence. It also seemed that differences in stress patterns of the individual words and differing syntactic relationships within declarative sentences would necessitate having a large inventory of intonation curves. These curves could probably be made dependent on the syntax in (something like) the same way that our intonational patterns are, i.e., there could be generating functions for the families of curves corresponding to the possible families of syntactic structures (as represented, for example, by phrase structure rules). These generating functions would account for the infinite generativity of the English language. Our proposal embodies a generative schema for assigning actual pitch levels to the words of a sentence which is sensitive to the syntax. Wherein lies the difference?

On the theoretical linguistic side, we have the same argument, put forward above, that the various links of our theory to other phenomena (tone languages, music theory) make it more interesting than the other theories which are only suitable for prosodics. Even from a more applied viewpoint, we have seen how wide a variety of data is analyzable under this theoretical viewpoint.

III.4 Some simple questions

It is generally accepted that wh-questions (or "information" questions) have pitch contours similar to declaratives while yes/no questions have a different contour. Below, we give short sample analyses of questions in terms of our basic tone group method. We will analyze wh-questions with the tone group (M)HL as for declaratives, and, the yes/no question with (M)LH.

III.4.1 data

We present some examples of questions. Sentences 1 to 3 are wh- questions and sentence 4 is a yes/no question. All of these questions were recorded by the same speaker (the one who recorded the declaratives shown in Figure 20).

WORD	PEAK FREQUENCY		
1.	(a)	(b)	(c)
what	196	179	192
is	200	192	204
the	169	182	182
answer	196	200	172
2.			
what	167		
rule	222		
of	200		
inference	182		
was	114		
used	108		

WORD	PEAK FREQUENCY		
3.	(a)	(b)	(c)
what	196	189	200
rule	196	196	217
of	147	161	182
inference	169	164	161
was	118	120	120
used	172	169	185
to	164	156	164
infer	152	149	159
line	145	139	152
fifteen	152	147	152
4.			
doesn't	182		
that	189		
observation	189		
seem	156		
exceedingly	161		
appropriate	192		

Figure 22. Questions.

III.4.2 Some wh- questions

There are some small differences among the pitch contours of the three recordings of sentence 1, but the general impression is that a mid tone on 'what' leads up to the elaboration of the rest of the tone group on the rest of the sentence. From this pervasive mid tone on 'what' we conclude that (at least in this type of question) the 'wh-' word is a minor lexical item and therefore able to receive the mid tone of the tone group.

1.

[what [is [the answer]]]

(a) [196 [200 [169 196]]]

(b) [179 [192 [182 200]]]

(c) [192 [204 [182 172]]]

In recordings 1a and 1c, there is a downstep from 'is' (the verb) to the noun phrase 'the answer'. The high to low tone contour agrees with the pattern we saw in the short declaratives above. We must say in these cases that the 'is' is sufficiently important in these clauses to avoid receiving a mid tone. In 1b 'is' does receive a mid tone.

'The' upsteps to the head noun in 1a and 1b but not in 1c which may be related to the relative importances of the two words in the noun phrase. It is also possible that the word 'answer' is slightly stressed in 1b. If we lowered the pitch on 'answer' in 1b, it would show the same pitch pattern as 1a.

2.

[[what [rule [of inference]]] [was used]]

[[167 [222 [200 182]]] [114 108]]

The next example (sentence 2) again shows the upstep (mid to high) from 'what' to the rest of the phrase even though it is not adjoined to the sentence but only to the noun phrase. Notice the large step between the noun phrase 'what rule of inference' and the verb phrase 'was used'; It is roughly an octave (222 herz to 114 herz). This indicates an isolation of the two phrases. Furthermore, the word 'of', which would normally upstep, does not in this case, indicating that the weights of different major and minor lexical items differ in isolated (stressed) phrases. Further evidence that this is a stressed phrase is the fact that 222 herz is above the normal pitch range observed for this speaker.

We have provided one level of syntactic simplification for example 3 because of its length and complexity. This simplification portrays the declarative fall pattern more clearly. Again, in 3b and 3c (and arguably in 3a), 'what' has a mid tone relative to the rest of the phrase. The interval of upstep is smaller here than in sentence 2, since the entire noun phrase is less prominent in the sentence; also it is not particularly isolated in relation to the rest of the pitch contour.

```
3.
  [ [what [rule [of inference]]][[was used] [to infer [line fifteen]]]].

(a):
  [ [ 196 [ 196 [147  169    ]]][[118 172]  [164  152 [145    152 ]]]].
  [ [ 196 [ 196     169       ]][     172   [164  152       152    ]]].

(b):
  [ [ 189 [ 196 [161  164    ]]][[120 169]  [156  149 [139    147 ]]]].
  [ [ 189 [ 196     164       ]][     169   [156  149       147    ]]].

(c):
  [ [ 200 [ 217 [182  161    ]]][[120 185]  [164  159 [152    152 ]]]].
  [ [ 200 [ 217     182       ]][[    185   [164  159       152    ]]].
```

The exact structure of the predicate ('was used to infer line fifteen') is not critical to this analysis, since an alternative structure:

[was used to infer [line fifteen]]

is also adequate for a description of the pitch contour in our tone group terms. We would like to point out that if the word 'to' had received a mid tone, as we might expect with the first syntactic structure, the second structure would not be adequate. As the facts lie, we need to 'explain away' the non-mid tone on 'to'. In sentence 3, as opposed to sentence 2, we do see examples of an upstep from 'of' to 'inference' (3a, 3c). We can relate this upstep to the relative neutral contour which the entire phrase takes on.

III.4.3 A yes/no question

Suppose that we were forced to choose a tone group for yes/no questions based on the sentence presented below. Let us suppose further that the choice was between (M)HL, 'declarative fall', and (M)LH[9].

4.

[doesn't [[that observation] [seem [exceedingly appropriate]]]

[182 [[189 189] [156 [161 192]]]
[182 [189 [156 192]]

The choice seems fairly easy to make: the (M)HL tone group doesn't fit even approximately to this sentence. We could say that 'doesn't', 'seem', and 'exceedingly' all upstep (mid to high) to the heads of their respective phrases. While 'doesn't' is believable as a minor lexical item, and 'seem' could be argued to be some sort of copula, it is hard to see what argument can be made for 'exceedingly'. If anything, we would expect it to have an exaggerated high tone.

---

[9] In reality, there are a great many more potential choices available, especially if some theoretical devices that have not been utilized in this research, e.g., boundary tones (Liberman, Goldsmith), are incorporated.

On the other hand, the (M)LH tone group is easily applied to this sentence. The predicate, [seem exceedingly appropriate], shows a consistent rising contour which corresponds to the low to high tone elaboration. There is no particular evidence that 'doesn't' receives a mid tone in this sentence since it may be part of the elaboration of the low to high contour for the structure AUX1 + NP + PRED and not (as shown here) part of a structure of AUX1 + CLAUSE.

We have not done enough further work in this area to make any definitive statements about the structure of question intonation, but it seems from this example that the method of tone group elaboration is one reasonable approach to empirical investigation there as well as in declaratives.

III.5 Phonology and the Real World of Phonetics

We have not yet attempted to specify exactly how the tone levels are translated into phonetic terms, what the assignment and smoothing procedures introduced above, Section II.1.2, are. We will attempt to make this specification exact below. We must also provide the mechanism to go from these tone levels to fully, phonetically specified intonation contours.

III.5.1 The size of the step

The first point, which is a fairly obvious one, is that the size of the increment (downstep or upstep) between any two words is a function of the pitch of the utterance. We would naturally expect that a person who speaks with a higher fundamental frequency will have larger intervals between the words in a sentence, ceteris paribus, than another person who speaks with a lower fundamental. This observation corresponds to the fact that equal frequency differences are perceived as smaller intervals when the interval occurs at a higher base pitch. We should therefore postulate some (musical) interval as being our basic step. Current suggestions[10] are that the interval of a third in musical terms seems to be a very natural one. Since there are twelve semi-tones in an octave, and each octave represents a doubling of the fundamental frequency, a minor third, consisting of three semi-tones, is represented mathematically by a multiplicative factor of 1.19 (= the fourth root of 2) and a major third, composed of four semi-tones, is represented by a factor of 1.26 (= the cube root of 2)[11]. At a frequency of about 200 herz the major third downsteps would be: from 200 herz to 159 herz to 126 herz to 100 herz (the octave). For minor third downsteps we have: from 200 herz to 168 herz to 141 herz to 119 herz. In this range, for minor third downsteps, we have an increment of roughly 30 herz. This figure is also reasonable in light of the data we have collected. Above, we saw several sentences in which the step at the least elaborated level of analysis (e.g., between clauses) was about 30 herz starting at 200 herz. We have used this basic step size for the sentences we have generated. While we would be more accurate to use the actual multiplicative factor (say the one for a minor third), we have simplified the system somewhat to use the step size for the initial frequency instead.

---

[10] Cf. Liberman (1975) pp. 126-128).

[11] The cube and fourth roots are explained as follows: Since there are four minor thirds in an octave (or three major thirds), each represents a multiplication of the frequency such that the total product is a doubling of the frequency.

There is an important second factor which influences the size of the increment. The syntactic location of the constituent, relative to the entire syntactic tree, determines what fraction of its 'natural' increment will actually be realized. By natural we mean the increment that would be realized considering solely the base frequency. Since the musical interval system operates on a power of 2 basis, i.e., normal harmonic relationships (in music) are related to each other as powers of 2 times their fundamental frequency, we have chosen to divide our basic interval by 2 each time we enter a lower constituent from its parent constituent on the tree. If we are considering a phrase which has no complex constituents, then we should see the paradigm downstep operating. If the same phrase is embedded in a larger phrase as one constituent we expect to see the same qualitative pattern but a different size increment (half the preceding increment) for the downstep. We could express this decrement in musical terms as a narrowing of the interval from a third to a second, and so on.

There is also a range of non-syntactic influences on the size of the interval. Additional emphasis or de-emphasis can be assigned to a word or constituent by increasing the step size. We incorporated this idea in our current synthesis program by using a step size twice as large as in the neutral instantiation for the downstep after a stressed word. Other non-syntactic interval size determinants include the speech style and location within the discourse[12]. In a speech style where each word in the text is somewhat isolated and distinct, or, at the introductory and concluding lines of a text the interval between tone levels will be larger than in neutral cases.

There are some predictions which result from making this quantitative specification of the increment size dependent on the syntactic bracketing.

1) We predict that the depth of embedding of a constituent makes a difference to the actual intervals that are used in the stepping of the intonation contour for that constituent. By putting this prediction in terms of the (musical) intervals involved rather than the increments involved we have controlled for the relationship of the increment to frequency.

2) Since we are claiming that a decrease in the step size occurs each time a new constituent is entered from a parent constituent, we expect that, unless there is a semantic interference, the step size

---

[12] More is said on discourse phenomena below, Chapter VI.

will eventually become zero. Our intuitions on the matter, based on sample casual tests, are that variations of several hertz are inconclusive in an intonational contour unless they affect the shape of the contour in terms of peaks and valleys. Thus at some level in the constituent structure, we will find that the intonation contour flattens out to the extent that there is no difference (other than random variation) between the frequency levels of the words in that constituent.

We can provide a rationale for this continual reduction of the interval for embedded constituents. It is easy to see that in a speech style where each word is emphasized, as in a dramatic reading, the intervals between words and constituents are fairly large throughout. Each word in such an utterance maintains an importance in the overall context. In normal speech, the more deeply embedded a constituent is, the less it contributes to the semantic interpretation of the uppermost parents. An extra prepositional phrase added to some constituent tends to make the head of the constituent less important to the "meaning" of the entire unit. As an obvious example, in the phrase "the man on the street", the constituent "the man" is less crucial to the interpretation of the entire phrase than would be the case if it constituted the entire phrase. The depth of constituent structure then is some measure of importance of the lower unit and (according to our rough hypothesis) is accordingly given a smaller stepping interval. While this notion does not constitute a full explanation of interval reduction as a function of syntactic embedding, the link between the psychological reality of language perception and interpretation, and, intonational facts does explain part of the relationship between structure and sound.

III.5.2 Words that downstep

Let us describe the smoothing function in some detail. After the tone levels are assigned to the words of the sentence, the actual contour connecting the different words needs to be imposed. This actual contour is assigned via the smoothing function which modifies the lexical contours of the words in a sentence. The smoothing function distinguishes between words in positions which mandate a retention of a contour similar to their lexical form and words which make a major modification of their contour. Phrase and sentence final words tend to retain close to their lexical contours as do emphasized words while phrase medial words do not.

A simple reason why there should be this difference in contours is the reductio argument that if not, the sentence medial words would sound as if they were in sentence final position. The primary objective of word medial smoothing is to insure that a smooth sentence contour is produced whereas the opposite effect is desired for stressed words, and phrase and sentence final words.

Comparisons of sentence medial occurrences with lexical contours has indicated that it is the end of a word that is typically modified when the word is used in a sentence. We will define the end of a word to be those segments of the word occurring after the maximum pitch, which is normally realized on the most prominent (stressed) syllable. For monosyllables, our definition of end yields an end that is almost the total length of the word. For polysyllables the end could be just the last voiced consonants if the word was stress final, or, it could be as long as the entire word if the word was stress initial. There are certain difficulties with this definition which we will discuss later.

III.5.2.1 Method one

When a word downsteps to the word that follows, there are several alternative modifications of the end of the (first) word. One possible modification that can be used is to level the pitch of the end from the point where it would have fallen below the (maximum) level of the next word. Using this modification we limit the fall at the end of the word to be just its lexical fall from the maximum pitch of the word until the maximum pitch for the following word is reached. We then give a level tone to the remainder of its (lexical) contour.

Let us give an abstract example: Suppose WORDA is assigned a peak frequency of 150 hz and WORDB is assigned a peak of 135 hz.; WORDA is 800 ms. long and its peak pitch occurs after 400 ms.; WORDB is 1000 ms. long and its peak pitch occurs after 800 ms.; WORDA falls in its lexical form from 150 hz to 110 hz, attaining a frequency of 135 hz 200 ms. after its maximum (600 ms. into the word). According to our leveling modification, WORDA would have its lexical contour for 600 ms. and then have a level tone of 135 hz. for its last 200 ms. WORDB would keep its lexical contour at least until its maximum (800 ms.), depending on whether it is sentence final or not.

In our experiments with this modification, we noticed that words so treated gave the impression of being sung. It seems that a (somewhat) long level tone will give that impression.

III.5.2.2 Method two

Another modification of the word contour we examined was to allow the fall at the end of the word to extend as far down as the pitch of the beginning of the word following. In our initial implementation of this method we used the first voiced segment of the following word as our low point and imposed a linear fall from the maximum of the (first) word through the end of the word to where it met the following word. While this method worked well for cases where the following word began with a voiced segment, our choice of using the first voiced segment in all cases sometimes produced odd results when the following word began with a voiceless consonant, since the pitch of the first voiced segment of the following word might be as high as the peak pitch for that word.

Our second implementation of the linear fall proposal was that in cases where the following word started with a voiceless consonant, we would define the length of the fall as some specified interval dependent on the maximum pitch of the (first) word. Thus, when the following word begins with a voiced sound we (smoothly) connect the two words by a linear fall. When the following word begins unvoiced, the end of the (first) word falls linearly to an interval below its own maximum. Based on observations of recorded sentences we chose an interval whose frequency was ten percent below the maximum for these cases. This method has yielded very reasonable sounding results.

We can illustrate this method by reconsidering the above example of WORDA and WORDB. We need to additionally assume that the beginning of WORDB is either voiced or unvoiced. Initially we will suppose that it is voiced and has an initial frequency of 120 hz. Our linear fall method will generate a contour for the end of WORDA that falls linearly from 150 hz to 120 hz over a time period of 400 ms., i.e., with a slope of .075 hz/ms for its voiced segments. If we suppose that WORDB begins with an unvoiced segment, then the end of WORDA will be a linear fall from 150 hz to 135 hz (=150 - 150*10%) for a slope of .0375 hz/ms.

One additional change to the linear fall smoothing algorithm has been made in the synthesis system. The change is that we have a dynamic redefinition of "the end of the word" to include as much as the entire word, even in cases where the pitch peak is not the initial pitch. Primarily in cases of upstep it is useful to make this redefinition.

III.5.2.3 Other methods

Other smoothing algorithms have also been implemented for limited experimentation. Two which deserve to be mentioned are "tonal smoothing" and "hyperbolic smoothing". "Tonal smoothing" is the imposition of one completely level pitch for the entire (voiced segments only) word. The effect of singing is quite pronounced in this case. "Hyperbolic smoothing" imposes a hyperbolic contour on the end of the word in connecting it to its "fall point". The equation determining the hyperbola is given by:

```
        PITCH (TIME) = ( A / TIME ) + K
where
        A = (max_pitch - fall_pitch) (fall_time) /  (fall_time-1)
        K = (max_pitch - fall_pitch * fall_time) /  (fall_time-1)
        TIME ranges from 1 to fall_time.
```

The net effect of using "hyperbolic smoothing" is to generate a quick, steep fall from the pitch peak, followed by a relatively level connection to the beginning of the following word. While perceptually, this method seemed good, it is our view that the only advantage for this smoothing over the linear fall is when a word is not stored correctly in the dictionary and de-emphasizing the very end of it improves the sound quality in general. Otherwise, there is no benefit for the "hyperbolic smoothing".

III.5.3 Words that upstep

Let us consider the different linear smoothing methods mentioned above in reference to a situation where a word steps up from its maximum to the maximum of the following word. With the "declarative fall" tone group, this upstep results from either a mid tone upstepping to the neighboring high tone or from the 'accidental' juxtaposition of tones from separate elaborations of the tone group. In some other tone group elaborations, the upstep might be more common than in the declarative fall.

According to the first smoothing function suggested, the contour for the end of the word should be made a level tone when its fall is below the maximum for the following word. Since, in an upstep, the following word has a maximum higher than that of the current word, the effect of this method would be to always level the entire end of the current word at its maximum pitch. We know from recorded sentences that this is not the case, and we therefore discard this method of word contour modification. We note that we have actually shown the inutility of this method only for upstepping sequences and only by implication for other sequences. Combined with the objectionable singing quality a sentence acquires in downstep sequences as a result of this smoothing algorithm, we discard it entirely.

The second suggested method fares very well in the case of upstep sequences. Since most words begin their contour somewhat below the maximum, prescribing a linear fall from the maximum pitch of the current word to the beginning of the following word still generally results in a falling contour. For example, suppose we had determined that in the phrase "the answer" the word "the" should be at a peak frequency of 124 hz while the peak frequency for "answer" should be 140 hz. The beginning of "answer" in such a case would be 112 hz and thus the fall from "the" to the beginning of "answer" would be smooth even though the phonemic sequence of words shows an upstep. This contour, in fact, duplicates those found in recorded utterances of the phrase "the answer". For cases where the following word does not begin with a voiced segment, the second implementation of the linear fall method has the advantage of preserving the fall regardless of the frequency of the first voiced segment of the following word.

III.5.4 Non-smooth smoothing

The linear fall algorithm discussed above gives sentence contours a connected realization that results in a smooth sounding synthesis. There are cases in normal speech which generate discontinuities in the sentential contour, despite regular tone group elaboration and assignment. These cases are: sentence and phrase final words, special emphasis and noun-noun compounds.

III.5.4.1 End of phrase and end of sentence.

We have already discussed the fact that at the end of a sentence, the last word is in almost its lexical form. Similarly, we can assign this type of contour elsewhere in the sentence. When we say "lexical form" we are really concerned with the fall at the end of the word. Typically, the fall for the last word in the sentence will be four times the interval of downstepping that occurs between the most major constituents of the sentence. The lexical reading of a word exhibits a fall of comparable magnitude. Assuming that the peak pitch of the word is about 200 herz, the lexical fall is about 100 herz, or four downstep intervals (about an octave, in musical terms).

With other constituent-final words we also observe an extended fall, which is roughly equal to three times the fall for the constituent. If a word is the last word in more than one constituent, as frequently happens in embedded constituents, the fall is based on the largest step, corresponding to the highest node which that word ends. Most constituent-final words will fall less than the sentence-final word, since the step size will be smaller in any embedded constituent than in the overall sentence.

For example, in sentence 1, discussed above, Section III.2.2, we analyzed the sentence as:

[[read the text] [if you can]].

We find that there is a fall on the word 'text' from its peak pitch of 167 herz to 92 herz (fall = 75 hz) and on the word 'can' there is a fall from 156 herz to 78 herz (fall = 78 hz). The sentence-final fall was only slightly larger than the phrase-final fall for the first phrase. The step between 'the' and 'text' in the above sentence was 25 herz (192-167), and, the step between 'you' and 'can' was 19 herz. These figures agree with our prediction that the sentence-final fall is four steps (156 hz is an octave higher than 78 hz) while the phrase final fall is three steps.

III.5.4.2 Special emphasis

In the discussion of 'emphatic stress' above (Section II.5) we pointed out the basic method of isolating a word by various pitch, rhythmic and loudness manipulations. A short experiment indicated that the relationship of the different manipulations could be ordered as producing most to least stress. Loudness or pitch isolation alone was not perceived as natural stress. Rhythmic isolation due to silences inserted around the word did produce stress. Combination of the silences, the additional isolation due to pitch manipulation and some loudening resulted in the perception of greater stress, and, additionally, lengthening of the word as well as pitch and rhythmic isolation was perceived as most stressed.

The pitch isolation was achieved by using non-neutral values for peak pitch assigned and for the fall assigned. Words following the stressed word are assigned tones with reference to the fall point of the stressed word and the syntactic structure. The peak pitch assigned to the stressed word is one step higher than the highest pitch in that constituent and the fall is one step in length which leaves the next word at the peak frequency for the constituent with the maximum step size in that constituent, i.e., it is as if the bracketing was re-arranged so that the stressed word initiated an isolated constituent in which it was followed by the rest of that phrase.

III.5.4.3 Noun-noun compounds

It is commonly accepted that the stress pattern in noun-noun compounds, such as "proof procedure", is different from what we would predict on the basis of adjective-noun sequences. For example, we have ambiguous phrases such as "steel warehouse", "English teacher", etc., which are only disambiguated between being noun-noun or adjective-noun compounds because of their stress pattern. We have found that in addition to modifying the durations appropriately[13] for noun-noun compounds we must modify the frequency contour for the second word in the sequence so that the second noun of the compound has a flatter pitch contour than would be predicted from its phrase final position. We can explain this in our system of tone groups and elaborations by attributing the prominence in this sort of phrase to the first noun. When we elaborate a tone group with a marked prominence we have an effect similar to stress. The high tone is elaborated to the prominent word and, only a relatively non-prominent contour is given to the other noun in the compound.

III.5.5 The end of a word

We mentioned above that there is a problem with the concept of the end of a word as being those segments of the word which occur after the maximum pitch. The obvious advantage of such a definition is in its computational simplicity. No knowledge about morphological structure need be taken into account in order to implement procedures based on this definition. In fact, this definition gives results that match the observed phenomena for monosyllable words with a lax vowel. Problems arise when we consider tense vowels which introduce diphthongization, usually accompanied by a pitch contour in the shape of a "peak and valley and peak", where the first peak (corresponding to the tensed vowel) is higher than the second (representing the off-glide). In polysyllabic words the same "peak, valley and peak" pattern can be found representing the stressed syllable and a following unstressed syllable. In such cases, it is only after the last peak that the linear fall should be used, otherwise the fall will wash out the diphthong (or unstressed syllable) pitch contour.

---

[13] Nouns are somewhat longer, in general than adjectives. see below, Chapter IV.

For example, suppose that we have three words, WORDA, WORDB, and WORDC, for which the pertinent information is given below:

	WORDA	WORDB	WORDC
pitch maximum (hz)	155	155	155
duration (ms)	800	800	800
time to maximum pitch (ms)	200	200	200
secondary pitch maximum (hz)	--	150	150,145
time to second maximum (ms)	--	600	400,600
lowest pitch before secondary maximum (hz)	--	135	145,135
calculated final pitch (hz)	135	135	135

Each word has the same maximum peak, but WORDA falls steadily whereas WORDB has a secondary pitch maximum 400 ms. after the first, and WORDC has two secondary maxima, one 200 ms. after the first maximum and the other 200 ms. later. WORDA is a monosyllabic word with a lax vowel, WORDB could be either a diphthongized monosyllable or a bisyllabic word. WORDC is either trisyllabic or has (at least) one diphthong. The linear fall proposal based on the simple definition of "end of a word" would produce isomorphic contours for each of the words of this example. Since each word has the same pitch peak at the same time and falls to the same final pitch, the slope of the fall would be calculated to be exactly the same for each word. The results obtained in this way would be essentially correct for WORDA and increasingly wrong for WORDB and WORDC.

If we used the same linear fall mechanism but applied it to our more complex definition of "end", we would achieve different contours for each of the above words. For WORDA the contour would be the same as under the old definition since it has only one pitch peak. WORDB would have a different contour under this proposal. Instead of leaving only the first 200 ms. of the lexical contour untouched, 400 ms. would be left as is. Then the slope of the linear fall would be calculated as going from 150 hz to 135 hz (instead of from 155 hz to 135 hz) over a time interval of 400 ms. (instead of 600 ms.). The treatment of WORDC is similar. The linear fall for WORDC is from 145 hz to 135 hz over a time interval of 200 ms. and the fall starts 600 ms. into the word. These contours are closer to the recorded ones than would be generated by the simpler definition of "end".

In order to implement this new definition we have to know, before processing a word, how many peaks there are in the word after its maximum. This can be done by including a number in the lexical entry for each word, taking nonnegative values. A value of zero indicates a simple curve for the word with no peaks following the maximum pitch. A positive value indicates the number of pitch peaks following the maximum, but does not specify the locations of the peaks. Alternatively, we could mark each subsidiary peak in a word and complicate the algorithm to utilize the extra information[14].

Our experiments indicate that the difference between the more accurate definition of "end" and the computationally simpler one we have suggested when applied to our linear fall smoothing procedure is not noticeably significant. The greater simplicity of the current definition of "end" is the reason it was implemented.

III.6 Other ideas that have been examined

During the course of our experiments, we have had occasion to examine several alternate proposals for generating intonation contours. We have not conclusively pursued them, but we have not found them promising and pertinent to our endeavors.

III.6.1 Pitch average vs maximum pitch

We stated an initial assumption above (Section II.3.1) that words could be considered as having phonological tone, which was defined to be any maximum pitch of the word. In fact, the bulk of our discussion up to this point has been in terms of the maximum pitch of each word. One alternative assignment for the phonemic tone level of a word might be the average pitch in that word. It would be no more difficult, computationally or conceptually, to use the average pitch across a word than it is to use the maximum pitch. In an idealized situation in which a word has a completely level pitch these two definitions coalesce. Empirical data might choose between these two definitions of the tone level because the pitch average definitions will cause the maximum pitch to be set higher than the pitch peak definition would. Perceptual experiments, testing whether the perceived pitches in different moving contours are the average, peak or minimum pitch of the contour, would be relevant to this question.

---

[14] If we mark the position in the lexical word (either directly onto the coefficients or by giving a time interval from the beginning of the word) where the fall is to start, and give two (often redundant) values for calculating the word contour: maximum pitch and pitch to fall from, we can make every definition of "end" computationally transparent.

III.6.2 Phrases, sentences and paragraphs

There has been a tacit assumption, for the purposes of this paper, to consider sentences as our intonational unit. We noted above in the discussion of the downstep that any nonelementary phrase exhibits the stepping properties. We also noted that within a sentence the constituent phrases exhibit downstep with respect to each other. The interclausal relationships that we have shown effectively disprove the claim that the phrase should be the effective unit for prosodic analysis. While it is a building block of sentential prosody (see Section II.4), it cannot be considered an independent unit. Its prosodic status (both pitch and durational) is determined by the larger sentence it is found in.

No larger scale analysis of spoken text (e.g., a paragraph) has been performed as part of this project, but we have generated a paragraph of text as a way of subjectively testing our prosodic rules. Several of our previous proposals[15] have been sampled by applying them to "the North Wind and the Sun" (see Appendix II). As a (somewhat unintentional) byproduct of generating this paragraph, we noticed two interacting facts that seem to bear on the prosodic relationship of sentences to each other in connected text. The first fact we noticed was that when phrases (smaller than a sentence) are generated so that each has the same contour, i.e., each starts and ends at identical frequencies, the effect on the listener is extremely tiresome. The second fact we noticed was that when entire sentences (or at least multi-clause fragments) are used, this effect of repetitiveness is considerably alleviated even though the same frequency range was used for all the sentences.

---

[15] Principally they are: different contour smoothing algorithms, pitch assignment algorithms, and duration assignment algorithms.

Another argument against considering paragraph prosody to be basic come from Nakatani (1975) where it was shown that a sentence recorded in isolation could be inserted into a recorded paragraph in place of one of the sentences there and no difference between the two versions of the paragraph would be noticed, at least under auditorium conditions. If this holds true under further examination, then it is further proof that sentence scale is more basic than paragraph scale for most prosodic facts. A discussion of the interactions of paragraph scope information with sentential prosodics is found below, in Chapter VI.

In the context of paragraph scale prosody we would like to relate a fact about tone languages that may be of some interest. When a series of tones in an utterance leads the speaker down out of his normal pitch range, it is very common for the speaker to break the sequence of tones, change to a new, high pitch register and then continue his melody from this new basepoint[16]. This suggests that one feature of paragraph intonation could be a similar sudden rise when some phrase or word would fall below the speaker's minimum pitch level, followed by a continuation of the contour as before.

III.6.3 Least squares fitting of intonational contour

One other method of determining intonational contours, at least on a preliminary basis, that seemed promising was using a least squares best fit to a recorded sentence by the concatenation of the individual words appearing in the sentence. From a purely formal point of view, this attack on the question of intonation looked very hopeful. An analytic solution of the linear equations determining the relationship between the recorded and concatenated versions of the sentence could be obtained, which would give us a 2n-1 (where n is the number of words in the sentence) degree polynomial which would serve as the (additive) baseline upon which the (scaled) words could be concatenated.

Assuming that the resultant curve should be a good approximation to the recorded sentence, we obtained the following equations. First, we have the equations for the generated sentential intonation contour.

---

[16] Leben, personal communication.

$$F_i(t) = C(t) + k_i \times (G_i(t) - Av_i)$$

$$C(t) = \sum_{h=0}^{h=2n-1} a_h \times t^h$$

where $f_i(t)$ is the recorded curve for word $_i$ at time t,

$F_i(t)$ is the resultant curve for word $_i$ at time t,

$G_i(t)$ is the lexical curve for word $_i$ at time t,

$C(t)$ is the (to be discovered) baseline curve at time t,

Av_i is the average pitch for word $_i$ at time t,

a_h are the (2n-1) coefficients to be discovered,

k_i are the (n) scaling factors to be discovered,

n is the number of words in the sentence, and

the standard error of the fit of the resultant (predicted) intonation curve, F(t), to the recorded sentential intonation curve, f(t), is given by:

$$\text{Total-Error} = \sum_{i=1}^{i=n} (\text{Error}_i)$$

$$\text{Error}_i = \sum_{t=tbi}^{t=tfi} (F_i(t) - f_i(t))^2$$

where tbi = time at the beginning of word $_i$,

tfi = time at the end of word $_i$.

To minimize the error we set the (partial) derivative of the total error with respect to each of the coefficients equal to 0. Then for each h, h = 1,...,2n-1, and each i, i = 1,...,n, we get:

$$p_{i,h} - \sum_{j=0}^{j=2n-1} (a_j \times q_{i,h,j}) - k_i \times r_{h,i} = 0$$

$$\text{where } p_{i,h} = \sum_{t=tbi}^{t=tfi} (f_i(t) \times t^h)$$

$$q_{i,h,j} = \sum_{t=tbi}^{t=tfi} (t^{h+j})$$

$$r_{h,i} = \sum_{t=tbi}^{t=tfi} ((G_i(t) - Av_i) \times t^h)$$

By taking times, t, across the entire sentence, instead of word by word, we can simplify this equation to

$$p_h = \sum_{j=0}^{j=2n-1} (a_j \times q_{j,h}) + \sum_{i=0}^{i=n} (k_i \times r_{h,i}).$$

We can form a (2n-1 long) vector of the p's, a (2n-1 by 2n-1) matrix of the q's, a (n by 2n-1) matrix of the r's, and a (2n-1 long) vector of the a's.

By a similar argument, and taking derivatives with respect to the scaling factors, we can arrive at the equation:

$$c_i = \sum_{j=0}^{j=2n-1} (a_j \times d_{i,j}) + k_i \times e_i$$

$$\text{where } c_i = \sum_{t=tbi}^{t=tfi} (f_i(t) \times (G_i(t) - Av_i))$$

$$d_{i,j} = \sum_{t=tbi}^{t=tfi} (t^j \times (G_i(t) - Av_i))$$

$$e_i = \sum_{t=tbi}^{t=tfi} (G_i(t) - Av_i)^2$$

We can put the (n long) c's and the (2n-1 long) p's into one (3n-1 long) vector ($VECTOR_1$). The matrices $d_{i,j}$ and $e_{i,i}$[17] will form the bottom of the (3n-1 by 3n-1) matrix whose top is formed by the q's and r's. The vector of the a's can be combined with that from the k's into a (3n-1 long) vector ($VECTOR_2$). With these manipulations, we can rewrite our best fit equations into a standard linear equation of the form:

$$VECTOR_1 = MATRIX \times VECTOR_2$$

which we can solve by inverting the matrix and multiplying:

$$VECTOR_2 = MATRIX^{-1} \times VECTOR_1$$

The resulting matrix of coefficients will be guaranteed, for realistic situations, to have an inverse. This means that there would be a solution for the value of VECTOR2 as a result of the least squares fit. VECTOR2 would provide the coefficients of the polynomial base curve as well as the scaling factors for the individual words.

---

[17] $e_{i,i}$ is the matrix whose diagonal is e_i and which has zeroes everywhere else, i.e., $e_{i,j} = e_i$ if $i=j$; $e_{i,j} = 0$ otherwise.

We found one technical difficulty in implementing this approach which was that the solution was very dependent on how the recorded sentence was segmented into words. Since the segmenting had to be done independently of the rest of the fitting procedure, it became a difficult task. Furthermore, we had to take into account pauses, matching of unvoiced segments, and other nonintonational aspects of the recorded utterance in order to make the recorded curve compatible with the concatenation lexical words.

A more serious objection to using this procedure is that it is not exceedingly illuminating. Even when the base polynomial is calculated and generated, it is not obvious how this polynomial curve relates to some general notion of prosodic features. This is especially true in light of the discussion above about pitch average and peak pitch as representing the phonological level of tone[18]. It is also conceivable that a mathematically best fit curve might undergo severe, although not linguistically significant, contortions in order to handle some of the phenomena of word to word smoothing that we have examined above (Chapter III.5 and especially Section III.5.2). Even granting this problem, the entire least squares fit approach does not deal with any of the phonological constructs which seem, on our analysis, to offer hope of a relatively simple solution to the question of English intonation.

---

[18] Even if we had revised the above equations to use the peak pitch instead of the average, the essential point of this argument remains valid. The most we could hope for from the best fit would be to obtain a series of curves from which we might, subjectively, be able to abstract the general intonation facts we are seeking.

Chapter IV

Metrics and duration

Most of the preceding discussion has centered on the pitch contour of an utterance as being the most important aspect of its prosody. In doing so, we feel supported by the view expressed in Bolinger (1965, p.95):

> There is also a virtue in the courage and good sense to call pitch pitch, instead of compressing pitches, intensities, and durations into a prosodic souse, cutting it into four sizes, and marketing it as large, extra-large, jumbo, and colossal "stress".

We said that what is commonly called stress is actually a complex result of the prosodic system that has a systematic perceptual basis but no direct acoustic correlate. We then proceeded to describe some interactions of "stress" with the pitch contour, while attributing other "stress" perceptions to duration and loudness modifications. We have largely disregarded loudness (intensities) contours since researchers have generally agreed[1] that loudness is the least significant of the factors involved in perception of stress. Our own examinations have given additional confirmation of this view.

We are left at this point with the question of duration assignment, both for cases of "stress" and non-"stress" utterances. Actually, by saying duration, we have already simplified the prosodic structure of an utterance since the rhythm (for example: the timing of heavy or stressed syllables) of the utterance is a more complex issue than measuring individual word durations. We have adopted a fairly primitive approach which attempts to approximate rhythmic effects by controlling word durations, with some reference to the syntactic structure and explicit stress markings. We are not claiming that our metric system will account for the intricacies (the "interesting" cases) of English rhythm, but we do think that it is fairly accurate for the majority of simple cases.

---

[1] Cf. Bolinger (1965), Fry (1958) and Gaitenby (1965,1975).

The quantitative duration information presented below is subject to errors stemming from the difficulty of segmenting an utterance accurately. It is often very difficult to tell (aurally and by using quantized pitch and loudness contours) where one word stops and another begins, especially when there are phonetic processes obscuring the boundary as in the assimilation of nasal sounds (e.g., "one of which"). Another source of segmenting error comes in the difficulty of separating pause time (the length of silences between words) from stop closures (both voiced and unvoiced) word initially, and unvoiced final frications. Thus a word could be given a duration twice as long in one utterance as in another simply because the judgements of where the words actually began and ended were different in the different utterances (or even in the same utterance). An example of the uncertainty associated with difficulties of measurement is the word "the", which shows a variation in length of up to three times the shortest measured length. A further difficulty comes from the fact that a given word may be pronounced differently in different utterances, even when the utterances are repetitions of the same words (with the same meaning). While some of this difference may be attributable to the metric effects to be discussed below[2], a fair amount of this variation is simply uncertainty in measurement.

IV.1 Some elementary facts

We mentioned briefly, in the introductory chapter, some of the linguistic factors which are involved in duration. They included word-level phonetic effects, syntactic and semantic phrasal effects and discourse effects. We can view these effects as hierarchically arranged, that is, the phonetic effects establish an isolation duration for a given word, the syntactic effects act on this isolation duration to yield a phrasal duration, and the discourse effects act on this phrasal duration to yield a final, actual duration.

As an example of the word-level phonetic effects, we can examine some word groups, composed of minimally different phoneme sequences. The four words, 'close'(/kloz/), 'nose', 'note', and 'cloak' are a good case. The difference between 'close' and 'cloak' is the substitution of a voiced final fricative for an unvoiced final stop. 'Nose' and 'note' also differ from each other in these distinctive features. The correspondence is not exact, since /k/ and /t/ have different points of articulation, but the similarity is substantial. We can also pair 'close' with 'nose' and 'cloak' with

---

[2] See the discussion of the word 'two', Section IV.3.

'note'. Here the distinctive difference is between oral and nasal stops, word initially[3]. The duration[4] of 'close' is 624 ms, that of 'cloak' is 549 ms, 'nose' is 569 ms, and 'note' is 484 ms. The "fricative-stop" difference is 75 ms for the /k/-initial words, and, 85 ms for the /n/-initial words. Focusing on the initial phoneme difference we see an "oral-nasal" stop difference of 73 ms for the final-fricative words, and a difference of 65 ms for the stop-final words. While further examinations of such pairs would be required to reach a firm conclusion[5], these differences agree with the general facts that fricatives are longer than stops (in general) and that oral stops are longer than nasals.

We can also see the effect of morpheme concatenation on word duration by a similar examination. Besides the cloak/nose words, Figure 23 also shows the decomposition of 'phonograph' as /phono/ + /graph/, and 'phonological' as /phono/ + /logical/. There we can see that the morpheme /phono/[6] contributes differently to the durations in each case although the difference between the two *phono's may be due to error in the recording. Notice that

---

[3] There is also the difference between a stop and liquid cluster, /kl/, and a single stop consonant. in addition, the point of articulation differs for /n/ is a dental nasal, while /k/ is a velar oral, but the principal distinction is oral vs. nasal.

[4] Durations in this section refer to recorded isolation durations, drawn from our lexicon.

[5] Other researchers have and continue to study this type of contrast. These contrasts have not been a focal point of this research and we mention them only in passing.

[6] We label the constructed morpheme duration with an asterisk.

WORD	LENGTH (ms)		
close	642	/z/-/k/	
cloak	549	107	/kl/-/n/
			73,65
nose	569	/z/-/t/	
note	484	65	
phonograph	808	*PHONO- (length)	
graph	608	200	
phonological	955	*PHONO- (length)	
logical	680	275	
phonograph	808	/t/-/n/	
photograph	868	60	
whoever	510	observed (e= 45)	
who	293	*-ever <= 135	
ever	319	*who- <= 330	

Figure 23. Segmental durations for selected words.

'phonograph' and 'photograph', which differ only in the oral-nasal stop phoneme, differ in duration by 60 ms. While still not conclusive, this agrees well with the differences in cloak/note and close/nose, seen above.

As a final example of this type of comparison we look at the combination of /who/ + /ever/ to yield 'whoever'. The rightmost column of that part of the figure gives durations abstracted from a pitch and volume analysis of 'whoever'. From these numbers[7], it is clear that some duration reduction is going on in combining the morphemes. A possible hypothesis would be that durational shortening similar to the syntactically induced shortening is involved in morpheme concatenation. We have not pursued this enticing possibility.

---

[7] The error of 45 ms is the uncertainty as to the end of /who/ and the beginning of /ever/ which is continuously voiced, but has a 45 ms region between volume peaks corresponding to /u/ and /E/.

Several experiments (Lehiste, Olive and Streeter 1976, among others) have shown that duration information can be used to disambiguate different possible syntactic structures for an utterance. The key fact is that the final syllable in a phrase is usually lengthened from its phrase medial length. In the hierarchical view of durational effects, the phrase boundaries are seen as modifying the phonetically predicted duration for the syllable adjacent to the boundary. Thus Klatt (1976) gives a formula for vowel length that involves a minimal length for each vowel (in the language's inventory) and a proportionality constant that varies with the syntactic environment, number of syllables in the word and the stress/unstress quality of the vowel. He gives a similar formula for consonant length. Semantic importance, novelty or focus can result in lengthening from the "neutral" duration (or smaller shortening from a lexical duration).

Gaitenby (1965) found that speech style results in a difference in "tempo" but not in the relative durations of segments. "In general, slow speakers tend to be slow all along the line in their acoustic segments ...." (Gaitenby 1965, p. 3.6).

IV.1.1 Simple experiments

There are several ways of testing a hypothesis about duration modifications. The most straightforward involves segmenting a large number of different utterances and statistically comparing the observed durations on a word-by-word basis. Another test would involve generating sample sentences embodying the duration contrasts desired and having subjects judge the contrasts. A third test is to generate durations for an utterance (phrase or sentence) and compare that statistically with an observed (spoken) duration.

The first two tests face error from the fact that duration contrasts do not exist in isolation. Pitch (and volume) contours interact with the duration contrasts, creating seeming length differences where none exist in the acoustic signal, and negating the perceptual effect of others. The durations from the first test suffers from the possibility of errors in segmenting the utterances. The third test is liable to "washout"; whatever contrasts may actually exist in the signal can be washed out as a result of the accumulation of these differences canceling each other in the average.

We have used the first sort of test to arrive at an estimate of the necessary duration modifications (See below). Other researchers (Huggins, 1972) have conducted serious tests on the second model. We have used informal listening tests on this model as well to ascertain the reliability of our predictions. Our testing of the third type was the most disappointing of all. We generated durations for about 250 utterances, ranging from one to fourteen words in length (maximum of 4.5 seconds long). The recordings to which we compared our generated durations were made by the same speaker who recorded our vocabulary. The recordings were made independently of this experiment, for use in the computer-instruction course in logic at Stanford University (1976). The experiment measured the correlation of the recorded utterance's duration to (a) our duration predictions and (b) to the sum of the lexical durations of the words from that utterance. The disappointment in the experiment was that a standard statistical regression for a linear relationship in both cases yielded correlations that were statistically significant ($p<.001$). The correlation for the generated durations indicated that correct predictions were made in most cases (the regression line had a slope of almost 1), while the summed lexical durations predicted a too high value of about 1.7 times the observed durations. The strong correlations of both predictions shows the "washout" effect -- negating (in both cases) any useful information that might be present in the results.

IV.2 Possible duration assignments

Perhaps we should apply the methodology that we used for pitch contours to rhythmic and durational studies. Then we would expect to be able to isolate recurring "rhythm groups", possibly equivalent to the classical meters, dactyls, spondees, trochees, etc. These rhythm groups would be elaborated and applied to syntactically analyzed phrases of English. This position is fairly close to that taken by Liberman (1975).

IV.2.1 Liberman's metric system

M. Liberman (1975, and, with Prince, forthcoming) has proposed an approach to rhythm in English that utilizes the concepts of a binary contrast, strong vs weak, and of a hierarchy of these contrasts to account for the relationships of stresses that are normally described by a number from 1 to 4 (most to least stress). His system, while imposing a good deal of formal linguistic machinery is conceptually fairly simple and it manages to account well for common facts about English rhythm.

IV.2.1.1 Description

The basic construct of this system is the strong/weak dichotomy. This should not be confused with the notion of stress/unstress syllable. According to Liberman, a weak syllable can be either a stressed or unstressed one, but a strong syllable must be a stressed syllable. Also, strong and weak apply to groups of syllables, as we will see below. By stressed (and unstressed) syllable we refer to the conventional definitions as used in The Sound Pattern of English: stress syllables are strong syllables (having consonant clusters or tense vowels), others are unstressed. Stress is a paradigmatic feature, ranging over substitutive alternatives, while strong/weak is a syntagmatic feature, being defined only for sequential contrasts. The assignment of strong and weak features is governed by the lexical assignment of [+stress], the structural bracketing, and several principles. Different rules are involved at the word level than at the phrase level.

The strong/weak difference is essentially a local one. Contiguous words contrast with each other. Sequential phrase nodes contrast with each other and with nodes dominating only words. All nodes (both lexical and phrasal) must contrast with something. For example, in the phrase "rather fancy madison avenue address" we assign the following structure (from Liberman (1975), p.198):

```
               R
           /         \
[     w                   s      ]
     /\                 /    \
    /  \     [      w         s  ]
   /    \         /   \       |
[  w     s ] [  w       s ]   |
   |     |      |       |     |
rather fancy madison avenue address
 +  -   +  -  + - -  + - +  -   +
```

The +'s and -'s are syllabic stress features and play no role in the assignment of strong and weak to the words. We assign strong and weak (s and w) to the subconstituents according to Liberman's version of the Nuclear Stress Rule (NSR) (1975, p. 198):

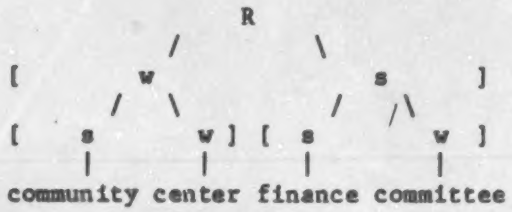
N —> s / [M __].

M and N represent arbitrary phrase nodes, i.e., [M N] is a phrase. Metric phrases do not cross over syntactic boundaries. The R at the top of the tree is simply a root designation; it performs no particular function in the metric system.

There is also a counterpart to the Compound Stress Rule (CSR) which is applied to word level rhythm (1975, p. 203):

N --> s / [__ M],

where M and N are as above. This would apply in word compounds such as "community center" and "finance committee" to produce [s w] sequences. We can see both of these metric rules apply in "community center finance committee":



There are also several complications on the metrical system outlined here to account for stress shifting (for example in the pronunciations of "thirteen" by itself and in the phrase "thirteen men"). Various sorts of special stresses and de-stressings express themselves as exceptions to the NSR and CSR, altering the pattern of weak/strong assignments.

The total hierarchical position of a word in a structure of this sort would determine its metrical quality. For Liberman most of the important contrasts are represented at the lowest level (the binary word contrasts), but the strongest word in a phrase will be the one that is the descendant through a chain of strong nodes from the root and the weakest word will be the descendant of a chain of weak nodes from the root.

IV.2.1.2 Translation into rhythm groups

Our proposal for tone groups as the fundamental construct of pitch contours has a natural counterpart in the expression of the basic stress rules (NSR, CSR) in Liberman's system as "rhythm groups". Instead of discussing "rules" that create structures of [s w] and [w s] at different phrasal nodes, we can name these two structures as:

iambs = [w s]
trochees = [s w]

These would be our basic "rhythm groups". They elaborate themselves by means of the metric tree which parallels the elaboration of our "tone groups".

IV.2.2 Some problems of implementation faced by Liberman's theory

What are some of the difficulties facing implementation of Liberman's metric theory into a speech synthesis program? The fundamental question is about exactly how the rhythmic information translates into timing parameters. Supposing that we had some algorithm for assigning strong and weak "beats" for English utterances, what is the link from those beats to word and syllable durations. Do we need to be able to put a specifiable syllable in a certain rhythmic position (at some particular time interval) in the utterance or is it sufficient that the word containing that syllable has the correct stress pattern and fits in rhythmically with the rest of the utterance?

Another implementation difficulty is the assignment of the rhythmic structures to actual utterances. We would need to be able to decide when and whether the CSR or NSR apply in a given sentence. While it is not very difficult to use these rules in individual cases, the difficulty in an automatic procedure is akin to the difficulty of doing a syntactic analysis automatically.

Still another question for a rhythmic theory is how it handles some of the phonetic facts that have been reported. We are thinking here primarily of clause final lengthening. Will application of the NSR to a sentence and the resultant heaviest stress being assigned at (or near) the end of the clause require the durational lengthening that is so often observed? Liberman's is a phonological theory, and we need to see how it can become a phonetic one.

In our synthesis system, we have an additional issue to decide. We have used a word concatenation system, based on a recorded lexicon. The rhythmic structures seem to use both words and syllables. To what extent should we modify our system to handle syllables[8]?

IV.2.3 Absolute and lexically based durations

When controlling the durations of words, we have several options. We can speak of basic units derived either from absolute durations, or from lexically based durations for the words. With either of these choices we have a further choice of relative or fixed relationships between the words. Each of these choices (as will be seen below) really represents a continuum between the most simplistic version of each end point. To a large extent it is a matter of taste and computational intricacy where on each continuum the choices are made.

---

[8] The choice between words and syllables is also an issue for pitch contouring. We use words marked for what we hope is the prominent syllable nucleus but marking all syllables in a word is an alternative to the current implementation.

An absolute duration is some (empirically determined) time span selected as the basic length for all the words in the sentence. Lexically based durations use the length of the lexical (recorded) forms of the words to determine their lengths in a sentence. For example, we might decide, after some investigation, that the longest word in a typical sentence lasts for 800 milliseconds (ms). We could then use that figure to determine the length of each of the words in a given sentence. That would be using an absolute duration. We could complicate the definition so that it takes into account the differences between monosyllabic and polysyllabic words, stressed and unstressed syllables, and differing levels of word-internal stress. We could allow an absolute time span per syllable, possibly broken down as stressed and unstressed. The discussion, above, of the possible implementation of Liberman's metric system, would utilize absolute durations for each syllable (or different absolute durations for [+] and [-] syllables). We feel that this general approach is best suited to a synthesis system which includes morphophonemics as part of its procedures or which utilizes segmental (or syllabic) synthesis. Our system, being based on a dictionary of recorded words, will lean to the lexical system described below.

The lexically based system differs from the assignment of absolute durations in that we assume that in a uniform (neutral) context, such as a single speaker recording a dictionary, each word will be recorded with a duration that correlates well with the number of syllables, their relative stress vs. unstress quality, and other phonetic relationships between phonemes. While using lexically based durations is sensitive to whatever variations a speaker (specifically the person recording the lexicon) may undergo from day to day, we believe that it is a reasonable basis for duration assignment. An example of the regularity of our recorded lexicon was given above in the discussion of the compositionality of a word into its segments. The several words cited there were recorded over a three year period by a single speaker. We have also noticed some instances where we attribute the generation of an incorrect duration to the lexical irregularity of a particular word. We are planning further experiments to compare the recorded lengths of words in our dictionary with predictions of word length based on the principles discussed in the literature[9].

From the above descriptions of absolute and lexical assignments, we can see that in its most complicated form (where each segment is assigned a context-sensitive set of durations) the absolute assignment should make largely the same predictions as the lexical assignment operating at its most regular.

---

[9] Especially recent work of Klatt.

IV.2.4 Relative and fixed relationships

The other dimension of duration assignment, assuming that a choice has been made between absolute and lexical durations, is the choice between fixed and relative relationships. This amounts, in phonological terms, to having some system of stress reduction or subordination. A fixed relationship involves setting the duration of a word only as a function of its basic unit, without regard to the relationship of the duration of any other word in the sentence to its own basic unit.

For example, we could fix each word's duration to be exactly its basic unit. Then if our basic unit was an absolute one, having fixed the relationship to be 1/1 ("duration-assigned"/"duration-of-basic-unit") all the words in the sentence would be assigned exactly the same duration. If we used a fixed relationship (again 1/1) but with a lexically based basic unit, each word would be assigned its lexical duration. We could, of course, assign each word a different, fixed relationship based on its semantic/sentence functional value. We might assign nouns and main verbs a relationship of 6/8, prepositions and other function words 3/8, etc. Then there would be a variety of durations assigned either with the absolute unit or the lexical unit.

A relative relationship would presumably entail using some generative system of stress subordination akin to the "rhythm groups" discussed above. In such a system the duration that a word received would be dependent on its structural relationship to the sentence and thereby to the duration assignments made to the other words of the sentence. Phrase final lengthening is one simple example of a relative duration relationship. The nuclear stress rule and the compound rule from "The Sound Pattern of English", Liberman's system and our implementation of his system are others.

IV.3 Data

In the next several figures we display some of the data from which we calculated the fixed duration relationships which we use.

Figure 24 and Figure 25 give the observed durations (in ms.) of words appearing in the paragraph "The north wind and the sun" as read by three different speakers[10]. All three speakers are natives of the northeastern United States. EA is a woman, AL and SIL are men. We have arranged the words according to word class, as they would be labeled in our prosodic system. The class of 'ad-words' combines the classes of adjectives and adverbs. The column labeled LEX gives the lexical (isolation) recorded length of the word and under each of the other columns we show the durations excised from continuous utterances. We can see the extent of variation in different instances of the same word. 'the' varies from 60 ms to 210 ms for AL, from 40 ms to 200 ms for SIL, and from 60 ms to 130 ms for EA. There is no reason to suppose that most of the variation is due to syntactic or semantic effects although some of it may be due to vowel reduction

---

[10] The pitch information from these readings is given in Appendix II.

WORD	LEX	(al)	(sil)	(ea)
			(TIMES x 10 MS)	
Pronouns				
HE	56	12,10	9,13	13
HIM	48	8	15	15,29
ONE	39	18	35	-
THEY	53	15	8	11
WHO	41	29	8	-
WHOEVER	56	-	-	31
Determiners				
A	41	13,15	6,4	7,11
HIS	55	15,25,17	7,7,10	16,10,6
THE	51	6,7,7,7,10,10, 10,10,10,10,12, 12,13,14,14,15, 15,15,21,21	4,5,6,6,7, 7,7,8,8,8, 9,9,9,9,9, 10,11,12,20	6,7,7,7, 7,7,9,9, 11,12,13, 16
WHICH	48	19	27	15
Auxiliaries				
BE	55	12	12	23
COULD	53	23	29	10,35
DID	48	13	11	-
SHOULD	55	11	19	-
WAS	52	14,17,13	19,15,11	2,11,13
WERE	47	15	17	17
WOULD	51	-	-	15
Conjunctions				
AND	58	30,20,15,12,10,9	17,14,11,10	13,9
AS	54	16,16	3,11	17,15
BUT	48	13	13	9
SO	53	22	20	-
THAN	58	12	14	-
THAT	54	20,14	15,31	21,12
THEN	62	31,54	15,17	23,17
WHEN	45	9	11	8
Prepositions				
ABOUT	31	-	-	17
ALONG	83	52	35	58
AROUND	73	48	45	-
AT	39	15	12	-
IN	41	15,14	6	7
OF	48	24,13	14,14	7
OFF	45	20,25	24,24	37,25
OUT	37	13	18	23
TO	48	11	2	11
UP	35	19	37	-

Figure 24. Sample durations arranged by word class.

WORD	LEX	(al)	(sil)	(ea)
Verbs				
AGREED	686	300	459	
ADMIT	363	-	-	363
BLEW	566	280,612	229,325	347
CAME	573	177	233	204
CONFESS	850	528	577	-
CONSIDERED	846	362	414	-
DECIDED	723	-	-	441
DECLARED	724	-	-	419
DISPUTING	785	600	580	685
DREW	597	-	-	126
FOLD	653	269	253	-
GAVE	713	164	247	-
OBLIGED	819	476	515	758
MAKE	503	-	-	200
MAKING	621	284	313	-
SHONE	713	267	257	300
SUCCEEDED	838	646	574	-
TAKE	581	236	186	220
TOOK	477	163	242	171
WRAPPED	542	353	381	441
Nouns				
ANALYSIS	969	403	-	-
ATTEMPT	615	400	459	-
CLOAK	549	287,301,313,384	337,362,439,440	359,281,465,508
FIFTEEN	850	563	-	-
F'NDATIONS	1139	648	-	-
LINGUISTICS	979	910	-	-
SUN	466	246,447,334,192	379,362,312	309,281,319
TRAVELER	727	407,331,462,401	466,427,399,357	473,450,465
TWO	441	320,188	337	664
WIND	572	475,458,338,360, 387	297,291,256	333,250,450
Ad- words				
CLOSELY	693	367	503	-
HARD	573	273	243	300
FIRST	573	262	266	-
IMMEDIATELY	909	864	565	513
LAST	625	340	535	-
MORE	469	280,182	238	191
NORTH	556	388,380,360,350, 280	290,359,374 390	303,300,326
ONLY	609	-	-	281
OTHER	440	281	249	-
PHONETIC	747	403	-	-
PHONOL'CAL	1114	615	-	-
STRONGER	667	425,450,625	502,457,499	431,403,536
WARM	625	480	307	350
WARMLY	718	491	409	570

Figure 25. Sample durations arranged by word class.

(the word "the" can be pronounced with a full vowel /e/, with a schwa or without any identifiable voicing).

Somewhat more useful information is given in Figure 26, Figure 27, and Figure 28. The durations shown in the previous two figures have been averaged (for each speaker) and are displayed as a percentage of the lexical duration. We have also included data for another male speaker BT. This new column was abstracted from a set of 34 separate utterances composed of 15 different sentences[11]. Each sentence was read one, two or three times with at most two of the readings being sequential. The utterances ranged from four to twelve words (5 to 18 syllables) and from one to four seconds in total duration.

For each speaker, for each word class, we averaged the percentage durations to get the underlined numbers in each column. A dash in a column indicates that the word was not spoken by that person. In Figure 27 we give two averages for EA's nouns and BT's ad-words, as well as two alternative percentages for two of BT's nouns. The second word-class averages were calculated by eliminating one of the words from the average ('two' and 'stuck', respectively) because they skewed the otherwise consistent averages. The alternative percentages for 'rule' and 'two' (for BT) are the result of separating

---

[11] We have already seen some of these sentences in the pitch contour data.

WORD	LEX	(al)	(sil)	(ea)	(bt)
Pronouns		36	19	36	42
HE	557	20	20	24	-
HIM	482	17	31	45	-
ONE	386	47	9	-	41
THEY	529	28	16	20	-
WHO	409	70	20	-	-
WHOEVER	560	-	-	55	-
YOU	385	-	-	-	43
YOU'RE	190	-	-	-	43
Determiners		33	25	22	45
A	410	34	13	21	28
HIS	548	34	14	19	-
THE	511	23	17	18	29
THIS	716	-	-	-	42
WHAT	436	-	-	-	53
WHICH	479	40	57	31	74
Auxiliaries		28	33	33	42
BE	563	21	22	41	22
CAN	639	-	-	-	51
COULD	534	42	54	42	-
DID	481	27	23	-	-
DOESN'T	607	-	-	-	54
IS	480	-	-	-	39
MUST	586	-	-	-	51
SHOULD	549	19	34	-	-
WAS	518	28	29	17	44
WERE	468	32	36	36	-
WOULD	508	-	-	29	35
Conjunctions		34	27	25	37
AND	575	28	22	19	-
AS	541	30	13	30	-
BUT	478	27	27	18	-
IF	405	-	-	-	43
SO	532	41	37	-	-
THAN	576	21	24	-	-
THAT	536	32	42	31	27
THEN	615	69	26	33	42
WHEN	448	21	24	17	-

Figure 26. Percentage durations arranged by word class.

WORD	LEX	(al)	(sil)	(ea)	(bt)
Prepositions		45	43	44	32
ABOUT	307	-	-	55	-
ALONG	828	63	43	69	-
AROUND	729	66	62	-	-
AT	394	39	29	-	-
FROM	575	-	-	-	19
IN	414	35	14	17	-
OF	479	39	29	15	29
OFF	454	50	52	69	-
ON	386	-	-	-	51
OUT	370	35	49	62	-
TO	475	23	4	23	28
UP	349	54	105	-	-
Verbs		50	52	58	70
AGREED	686	44	59	-	-
ADMIT	363	-	-	100	-
BLEW	566	79	49	61	-
CAME	573	31	41	36	-
CHOOSE	623	-	-	-	56
CONFESS	850	62	68	-	-
CONSIDERED	846	43	49	-	-
DECIDED	723	-	-	61	-
DECLARED	724	-	-	58	-
DISPUTING	785	76	74	87	-
DREW	597	-	-	21	-
EXPECT	843	-	-	-	77
FOLD	653	41	39	-	-
FOUND	575	-	-	-	53
GAVE	713	23	35	-	-
OBLIGED	819	58	63	93	-
IMPLIES	857	-	-	-	63
INFER	610	-	-	-	73
MAKE	503	-	-	40	-
MAKING	621	46	50	-	-
PASS	615	-	-	-	53
READ	405	-	-	-	87
REQUIRES	808	-	-	-	80
SHINING	516	-	-	-	92
SHONE	713	37	36	42	-
SUCCEEDED	838	77	68	-	-
TAKE	581	41	32	38	-
TOOK	477	34	51	36	-
TYPE	424	-	-	-	87
USED	609	-	-	-	53
WRAPPED	542	65	70	81	-

Figure 27. Percentage durations arranged by word class.

WORD	LEX	(al)	(sil)	(ea)	(bt)
Nouns		64	67	82(65)	72
ANALYSIS	969	51	-	-	-
ANSWER	575	-	-	-	80
ANTECEDENT	887	-	-	-	83
ATTEMPT	615	65	75	-	-
CLOAK	549	58	72	73	-
CONDITIONAL	842	-	-	-	67
CONSEQ'NCE	1042	-	-	-	72
CONSEQUENT	875	-	-	-	78
FIFTEEN	850	66	-	-	67
F'NDATIONS	1139	57	-	-	-
INFERENCE	879	-	-	-	69
LINE	535	-	-	-	73
LINES	653	-	-	-	85
LINGUISTICS	979	93	-	-	-
RULE	373	-	-	-	83:94,74
SUN	466	65	75	65	53
TRAVELER	727	55	57	64	-
TWO	441	58	76	151	48:62,39
WIND	572	71	47	58	-
Ad- words		62	60	57	101(65)
CLOSELY	693	53	73	-	-
HARD	573	48	47	52	-
FIRST	573	46	46	-	-
IMMEDIATELY	909	95	62	56	-
LAST	625	54	86	-	-
LOGICAL	680	-	-	-	71
MORE	469	49	51	41	-
NORTH	556	63	64	56	-
ONLY	609	-	-	46	-
OTHER	440	64	57	-	-
PLEASE	712	-	-	-	58
PHONETIC	747	54	-	-	-
PHONOL'CAL	1114	55	-	-	-
STRONGER	667	75	73	68	-
STUCK	552	-	-	-	172
WARM	625	77	49	56	-
WARMLY	718	68	57	79	-

Figure 28. Percentage durations arranged by word class.

utterance final occurrences of the words from other instances of the words. In each case the larger percentage is the utterance-final value. The noticeable difference between the two percentages should be predictable from the fact of phrase final lengthening mentioned above.

If we expected the absolute duration scheme in its simplest form to be reasonable, we would look for coherences in absolute length (in a word class) regardless of what the lexical duration was. The first two figures are pertinent to this question. In the other three figures, we concentrate on the proportionality between lexical and in-use durations. In cases where all of the words in the class are of the same length (as measured in number of syllables), the difference between lexical and absolute durations will not be very apparent. If we look at the class of nouns, where we have a distribution of mono- and poly- syllabic words, we see that using percentages has leveled much of the variation in the absolute durations. For example, the words 'sun' and 'attempt' have the same percentage duration (65%) for AL and SIL whereas their absolute durations are clearly different. This is because the polysyllabic word is longer in its lexical form as well as in its utterance-medial duration than the monosyllable.

Notice that for all of the variation of the actual durations, of the percentages within each word class and of the percentages for a given word for different speakers, we find good general agreement in the word class percentages for the different speakers. The greatest disagreements occur in the class of pronouns and determiners, with the closest agreement for nouns and ad-words. Better agreement would probably result if lengthening had been explicitly taken into account. The BT nouns would have an average closer to the other speakers if the sentence final occurrences had been left out[12]. A similar observation holds for the verbs as well. We expect final lengthening to be more pronounced in general in the recordings from BT, since each was separate from the others, whereas the other three speakers recorded their text continuously without pause from beginning to end.

In Figure 29 we present some of our previously discussed sentences, this time together with duration information derived by segmenting (aurally) the recorded sentence and the fraction of each word's lexical duration as found in that recorded sentence. The line labeled (a) under each sentence gives the averaged (from three utterances) duration of the word in the sentence. The (b) line gives the percentage of lexical duration for that average value.

Sentence four seems hard to explain in our framework, but there are some aspects that we can handle. We have found a regularity (at least for the speaker BT) that he inserts 150 ms of silence before a stressed word. A word in quotes can be assumed to be stressed, so we can subtract 150 ms from the end of 'type' (which ends in a voiceless plosive and is therefore susceptible to duration error) for silence. The duration of 'type' is then only 84% of its lexical duration which is within the range of verbs that we have observed.

---

[12] Besides the two words shown in the figure, the words 'answer' and 'antecedent' only appeared in this corpus in final position.

1.	Please	type	the	answer.
(a)	419	315	122	470
(b)%	59	80	27	82

2.	The	rule	can	be	found	on	page	seven.
(a)	110	348	205	111	331	294	359	526
(b)%	29	89	28	32	53	51	57	86

3.	Read	the	text,	if	you	can.
(a)	352	168	653	166	129	421
(b)%	87	34	100	38	33	61

4.	If	your	stuck,	then	type	"hint".
(a)	182	189	924	249	508	317
(b)%	45	99	167	40	120	63

5.	Rule	aa	requires	two	lines,	one	of	which	must	be	a	conditional.
(a)	346	476	647	165	739/422	152	75	282	298	105	100	582
(b)%	93	—	80	38	113/64	41	16	66	51	19	24	69

6.	what	is	the	answer?
(a)	216	183	113	464
(b)%	49	38	25	81

7.	what	rule	of	inference	was	used	to	infer	line	fifteen?
(a)	270	267	157	653	214	324	120	443	395	532
(b)%	62	71	34	74	44	53	28	73	74	67

Figure 29. Durations of words for selected sentences.

The first three words of the sentence are longer than we would expect. It seems that there is some reason for the entire clause to be said slowly. But sentence three which also has a three word first clause is also somewhat slow (though not as slow as in sentence four). Possible explanations would be: (a) short sentences (and clauses) are said more slowly than longer similar syntactic units; or (b) the rhythm of short clauses (these have only three syllables) causes a slowing down of the rate of speech to accommodate the syllable differences. Possibility (b) could account for the difference between the durations in sentences three and four by the different pattern of stressed and unstressed syllables (3: + - +, 4: - - +)[13].

In sentence five, two averages and percentages are given for 'lines'. The first (larger) values give the length of 'lines' as including the silence following the word, the second values leave off the silence. Without the silence added, 'lines' has approximately the percentage duration that we associate with nouns. These samples are suggestive of several general possibilities for incorporating aspects of a relative duration assignment.

1) There is a tendency to lengthen the last word in an utterance.

2) Lengthening may retract if the word in final position is not a "stress bearing" word. We interpret "stress bearing" as equivalent to major lexical item, but this is only an approximation.

3) Words at the end of a major syntactic unit will also tend to lengthen.

4) The rhythm of stressed and unstressed syllables may affect the durations of words, especially in short phrases.

5) Silences are somewhat interchangeable with increased durations to give the percept of greater length.

---

[13] See the discussion above of a possible implementation of Liberman's assignments.

IV.4 Our current approximation to duration assignment

We have implemented a lexically based, largely fixed duration assignment system partially based on the data given above. The other basis for the assignments is the results of informal listening in a variety of contexts to determine the naturalness of different possibilities. We have assigned durational fractions to different parts of speech. Since the lexicon contains the parts of speech of the words for use by the syntactic and semantic components of the grammar in any case, we are not adding extra material to the lexicon by using this method. Our present system uses the distinctions in duration which are summarized in Figure 30, below.

PART OF SPEECH	DURATIONAL FRACTION
Noun, Verb	6/8 (75%)
Ad-word: Adjective, Adverb	5/8 (63%)
Preposition, Complementizer	4/8 (50%)
Determiner, Copula, Pronoun	3/8 (38%)

Figure 30. Duration assignments.

In addition to the above assignments, which represent the fixed relationship aspect of our duration control, we have some aspects of a relative relationship. We assign a fraction of 8/8 (=full) to the last noun, verb or ad-word in a sentence. If the last word in the sentence is not from one of these word classes, it is lengthened but not to its full lexical length[14]. Phrase final words are also lengthened.

---

[14] Adjectives, prepositions and complementizers receive 7/8 (88%) of their lexical length; prepositions, copulas and pronouns receive 6/8 (75%).

IV.5 Deficiencies and alternatives

IV.5.1 General comment

We regard our current duration assignment system as an approximation that has been reasonably successful in generating normal sounding utterances. One indication of how good an approximation we have made is seen in considering a study by Gaitenby (1965) that has produced evidence which seems to be in accord with our largely fixed lexical method of duration assignment. From the segmental analysis there of the recorded sentences:

1) Why don't you?
2) Why don't you get tickets?
3) Why don't you get tickets for tomorrow?
4) Why don't you get tickets for tomorrow night?
5) Why don't you get tickets for tomorrow night's program?

we can abstract the following duration chart for these words[15].

---

[15] What we call base duration is the duration of the word in the shortest sentence in which it appears. Medial duration is the duration in the longest sentence in which it appears. Words that were not sentence final for their base duration have their base duration and percentages enclosed in square brackets.

WORD	BASE DURATION	MEDIAL DURATION	PERCENTAGE (MD/BD*100)
why	[200]	90	[45]
don't	[480]	150	[30]
you	200	120	60
get	[260]	210	[45]
tickets	480	360	75
for	[80]	60	[75]
tomorrow	520	350	68
night('s)	340	250	75
program	—	610	—

Figure 31. Duration relationships abstracted from Gaitenby(1965)

In the above examples, it seems that the shortening of durations from their lexical form roughly follows our fixed lexical system. Also, the same example is evidence that only the differences between sentence final and sentence medial durations were significant. This finding is also in accord with our experimental conclusions here.

IV.5.2 The question of differential shortening

In our synthesis of sentences, the way we actually realize our duration assignment procedure, all segments of the words are shortened in proportion to their original duration. We mentioned above that final syllables are usually lengthened more in word lengthening contexts, and therefore they should be shortened most for word duration shortening. In addition, it has been noticed that unvoiced sounds become less distinct than is desirable when the entire word is shortened. Independently (Gaitenby 1965, Huggins 1972), it has been noticed that when speakers change the length of words, they do so by changing (in general) primarily the lengths of the voiced segments of the words, leaving the unvoiced segments at about their full (lexical) length. We have not tested this hypothesis ourselves, but it would not necessarily increase the computational load of our synthesis to perform this type of word shortening.

IV.5.2.1 A possible implementation of Liberman's system

We will propose here a sketch of an implementation of something akin to the rhythmic theory of Liberman[16]. We have not tried to actually incorporate the following program into our system, but it could be the basis for further improvements.

The first stage is the assignment of lexical stress patterns to individual words. This could be based on a phonetic decomposition of the words[17]. The stress pattern we are interested in is simply a string of +'s and -'s, representing [+stress] features, not the final subordinated stress pattern. We could decide the assignment of these stress features from strong and weak syllables as mentioned above. For example the words the, to, one, etc. are lexically marked with a single - from their [-stress] feature. The monosyllables I, gave, word, etc.are lexically marked with a single + because of their [+stress] nature. A more elaborate example:

su per cal i fra gi list ic ex pi al i do shus.
+ - + - + - + - + - + - + -

---

[16] The degree of similarity of our proposal to his is given by the formula OURS = (NOT) VERY^n MUCH LIKE HIS, where n is zero or greater and is chosen by the reader and the parenthesized 'not' is optional.

[17] Cf. Hunnicutt (1976)

Pi was assigned a - since its tense quality is derived from a rule tensing the first of two successive vowels and not from being in a strong syllable. We see here a (presumably normative) pattern of alternating +/- syllables. In words which do not exhibit this completed pattern, we modify the lexical pattern so that the word fits into the formula

$$(-)\ [+\ -]^{n}\ ,\quad n > 0 \text{ or } n = 0,$$

an optional - at the beginning of the word (an upbeat), followed by pairs of + and -. A word with only a - is fine, but one with only a + would need to be "completed" to + - with the insertion of a - at the end of the word. In cases of successive + or successive -, there are two possibilities. We could reverse the polarity of the second + or of the first -. Alternatively, we could insert additional +'s or -'s to ensure completion of the sequence. Presumably, inserted features will be added to the syllable they complete. Some sample applications of this pattern to words are given below, Figure 32. There the word "Pamela" can be completed in two ways. One way would insert a + between the two -'s, while the other would change the first - to a + and then insert a - between the two +'s. In "nominalizations" all that needs to be done is to change the second - to a +. With "pontificate" we can insert a - between the first two +'s or else change the second plus to a - and then add a + between the pair of -'s just created. With either method we would add a - at the end of the word.

```
   Pa me la                 Pa me la
    +  -  -                  +  -  -
    |  |                        |
    +- +                        -+

          nom i nal i za tion
           +  -  -  -  +   -
                 |
                 +

pon ti fi cate           pon ti fi cate
 +   +  -  +              +   +  -  +
     |     |              |         |
     -+    + -            +-        + -
```

Figure 32. Some samples of rearrangement of lexical stress patterns.

After we have assigned our completed lexical patterns, we perform the phrase level rhythmic manipulations. Within a given phrase (syntactic) structure (possibly marked for semantic prominence) we would add +'s as necessary to insure continued alternations of + and - patterns throughout the phrase. In addition, we subdivide the phrase into a hierarchy of weak and strong positions. We could force the +/- sequences to "agree" with the strong/weak hierarchies by adding [+ -] units to strong words and nodes (+'s and -'s not assigned to words could become silences), or by removing a + from a [+ -] combination for a weak word or node. While the leftover - is a viable upbeat in this system, it requires "completion", either with a + of silence or by lengthening of the preceding word. For example, clause final and prominence induced lengthenings would be accounted for by the addition of [+ -] sequences to those words. By ordering the insertion of +'s after subdivision of the phrases and insertion of +'s and -'s on the phrase basis, we could get different rhythmic predictions in some cases.

Finally, using the sequences that were assigned, we predict durations for the words and silences of the sentence[18]. For each [+ -] pair we predict N ms of length, where N is empirically determined, but independent of phonetic structure (an absolute duration as described below). An isolated - could be given either one-half or one-third of the length of a pair[19]. Each word would then be shrunk or lengthened to the summed duration for that word. If found to be desirable, we could perform this calculation on syllables rather than words. Otherwise the exact placement of the sequences within a word is not critical.

---

[18] The following account is to be read as a beginning hypothesis for further research. The work of Liberman, whose metric structures we have adapted, and Klatt indicate that there is more to duration and rhythm than we are about to propose.

[19] Some other fraction could probably also be justified, but we were thinking of the length of an upbeat in duple and triple meters.

Chapter V

The syntactic component of the model

There are two matters pertaining to our use of syntactic structure in the prosodic synthesis system that will be discussed here. One is simply a description of the rules that comprise the syntactic analyzer. The other is a more complicated issue, which underlies the whole linkage of the phonology/tonology with the syntax. Several variations of the issue are possible. One version of the question is how can/should we define (theoretically) and recover (in practice) the constituent structure of an utterance. There are both abstract and concrete criteria for the answers to both parts of the question.

An abstract answer to the theoretical question is an attempt to discover the "correct" structural descriptions of English utterances. The concrete side of the question is only concerned with being "correct" in relationship to the prosodic system; to the algorithms which map pitches (and rhythms) from the elaborated structures to the string of words. We saw above, Section II.6, in the discussion of contraction phenomena that a purely surface syntax, when used with our prosodic assignments, made some incorrect predictions. These missed predictions occurred where a more "correct" structural analysis would have enabled our prosodic system to correctly predict where contractions would occur. Since the actual prosodic system does not predict contractions, but only intonation contours, the incorrectness of the structures generated can be neglected in practice. They may be "wrong" but not enough to matter. On the other hand, if we adopted some other system for defining phrase structure the differences between the sets of syntactic structures might be critical for generating reasonable intonation contours. In that way there could be a "right" or "wrong" syntax, relative to our prosodics.

The question of recovering the syntactic structure automatically for practical use has its own abstract and concrete solutions. Can a simple construction of a phrase structure lead to usable prosodics? If some such simple construction were effective, is there a need to expend greater effort (computational or intellectual) on producing a more complex phrase structure? It has been hypothesized, for example, that relatively simple "local parsers" (Allen, 1968) which only provide phrase level structural information are adequate for sentence-level prosodic work. Similarly, our intonation system does not utilize phrase node labels and therefore the parser does not provide them. Also the parser is not part of a "grammar of English" in the sense that it does not produce grammaticality judgements about the strings of words which it parses. It merely tries (as best it can) to parse each string of words on the assumption that the string is grammatical (i.e., that it can receive a good parse according to the patterns in the parser). On a concrete plane there is the matter of constructing a parsing system that can achieve the type of syntactic analysis which is desirable on the more theoretical criteria.

V.1 Some complicated parsers

Our system for syntactic parsing is a surface syntax model of English but that is not an indication that some other model, such as a transformational grammar, would be unsuitable for expressing linguistic generalizations or for constructing a parsing algorithm. Since we will discuss our system in detail below, only a brief remark will be made here.

V.1.1 Our parser -- overview

Figure 33 gives an outline of our parser which uses linguistic patterns to decide how constituents are to be constructed and combined. This algorithm uses certain basic constructs (here: noun, verb and prepositional phrase; above: conjunctions, or, articles, auxiliaries and prepositions) to achieve a preliminary structure for the sentence. It then fills in the structure so established by creating more complex constituents. The final step in this parse procedure is to assign all as-yet unanalyzed words to some phrase. The basic motivation in this step is that English is a right-branching language, i.e., most of the complex constructions in English occur on some right branch of the syntactic tree. An example of this complex right-branching structure is a noun with a relative clause 'suspended' from it as in "John, who came home late last night." The structure for this phrase is

[John [who came home last night]],

1. Find the simplex noun phrases, verb phrases and prepositional phrases in the sentence.
2. Use these phrases, along with unphrased words, to form more complex phrases by looking for specified elements and then associating other constituents into the phrases.
   a. Some specified elements are searched for from the front of the sentence; some are searched for from the end.
   b. Associated constituents may be before or after the specified element.
3. Complete the structure by including any unphrased words as either (a) or (b).
   a. their own phrase, if there are enough words together.
   b. a left sister to some constituent, the created sisters to be dominated by some single node.

Figure 33. Overview of our surface parser.

one word adjoined to a clause on the right. While not every structure in English is right branching, this is a useful guess for the parser to make when it finds no other analysis.

Still more complicated parsing procedures are conceivable. We saw above in the discussion of stress reduction and auxiliary contractions (Section II.6) that a purely surface parser which utilizes only (linguistically motivated) structural information (no semantic information is used) will produce, for certain types of sentence, a parse that results in an incorrect prosodic assignment. What is particularly missing from surface parsers is the capability to deal with missing and moved constituents[1]. Let us consider a simple-phrase structure parser, which is non-recursive; there are no embedded constituents in its parsed structures. Such a parser is not subject to this difficulty to the same degree as our full surface parser since the limited structures available as simple phrases are unlikely to contain either moved or missing words that make important structural differences. We can compare the results of the surface and transformational parsers using the example from our previous discussion of stress reduction, which we repeat below, together with different possible parses. Parse (a) would come from the simple-phrase parser, (b) would come from a full surface parser, and, (c) could result from a transformational derivation.

---

[1] Missing constituents and discontinuous constituents were a major motivation (historically) for incorporating transformations and/or semantics into accounts of syntactic structure.

Eager though I am to win, I would never cheat.
(a) [Eager though] [I][am] [to win] [I][would never cheat].
(b) [Eager [though [I [am [to win]]]]] [I [would never cheat]].
(c) [Eager [though [I [am ___ [to win]]]]] [I [would never cheat]].

The focus of attention in this figure is on the different parsings for "am to win." The simple-phrase parse for those words shows two separate phrases. In terms of the prediction of contractions, this parse could agree with either of the other two, either incorrectly allowing contraction as in (b) because of a tonal assignment upstepping the word 'am' to the phrase "to win", or not, as in (c), instead, correctly assigning separate prominences to the word and the phrase. Parses (b) and (c) make conflicting predictions, with the empirical consequences supporting parse (c).

From a purely practical point of view, we note that the bulk of English sentence structures do not exhibit these deletions and movements. The additional complexity due to incorporating derivational schemes or semantics into a parser must be balanced against the need for correctness in these limited cases. While the surface parser does not always yield the best parse, it comes very close in a large number of cases.

V.1.2 More complicated parsers

Two parsers which perform more complicated syntactic analysis than our surface parser are those by Kaplan (1976) and Marcus (1974, 1976). Both of these attempt to produce structures which would make the correct predictions in the contraction cases cited above[2], that is, express generalizations about moved and deleted constituents. Neither parser tries to reproduce the derivational history of a sentence as part of a parse.

V.1.2.1 Kaplan's ATN parser

Kaplan's parser is an augmented transition network. A simple transition network is a graph of connections ('arcs') between possible states. In a parser each state represents some constituent which is recognized when the parser arrives at that state. The syntactic patterns which were the basis of our parser are represented in sequences of arcs and states of the ATN. Some sequences of states which would otherwise appear many times in the parser can be factored into subroutines which are separate transition networks. This transition network can be referenced by an arc between two states, as if it were a simple word recognition. In this case, the sub-network must give a successful recognition in order for the parse to proceed. These sub-networks normally represent phrase constructs like 'noun phrase', 'verb phrase', etc.

---

[2] Needless to say, this expression of their aims is totally our own.

The augmentation of the network comes from allowing the arcs to represent (unlimited) computations. These computations can be used to bracket or re-bracket constituents, to assign labels to various elements, and to perform tests besides the simple constituent recognition which the states represent. An important augmentation in Kaplan's system is the 'hold cell'. Some of the arcs have the possibility of placing a recognized constituent (either simple or complex and usually a noun phrase) into a designated cell, while other arcs are enabled to use this cell instead of performing a sequential sentence constituent recognition. Use of the hold cell allows this parser to analyze sentences with unbounded movement or deletion, which our purely surface parser could not, while avoiding reference to derivational histories. Unbounded leftward movement, which is the main example in English, is recognized by placing the (supposed) moved constituent into the hold cell[3] and then whenever a pattern may require a constituent of that type the hold cell is emptied of its contents.

---

[3] The supposition can be based on the absence of a syntactic pattern utilizing the constituent at that point.

An important difference between the organization of the ATN and that of our parser is that the ATN completes its parse of one (the current) constituent and then processes the next word or constituent in sequence. It is always building up a unified structural description. Our parser will create many independent constituents and then try to combine them into more complex constituents. The difference between these two approaches shows up with "garden path" sentences, such as "The horse raced past the barn fell." The ATN will initially follow a "garden path" in mis-analyzing a sentence until it can proceed no further. At that point it has to back up and recreate at least some of the parse, making some different choices from the structures assigned the first time. In the sample sentence it will parse "the horse raced past the barn" as a full sentence and then have no analysis for 'fell'. Then it will have to go back and analyze "raced past the barn" as a reduced relative clause in order to fit 'fell' into the structure of a declarative. Our parser will always maintain the integrity of the simplex structures created in the earliest part of the parse, even if higher level structures are difficult to form from these constituents. In the above sentence, assuming 'past' is marked (or recognized) as a preposition, the simple constituents "the horse", "past the barn", "raced" and "fell" will be parsed in the first run-through. The parser is then free to create any grouping of complex constituents it can. In fact our parser can only create a single structure for any sentence and thus cannot back up in case of parsing difficulties.

Another difference between the parsers is that an ungrammatical 'sentence' will require significantly more parsing than a grammatical one does from an ATN but not much more from our parser. This is because the ATN will attempt to recover from possible "garden paths" until all possibilities for parsing the sentence have been exhausted. Our parser will simply leave whatever constituents it can find sitting around and then give up. The ability to give up quickly is valuable in a setting where ungrammatical utterances may be encountered. The concomittant drawback is that our parser will tend to compound its errors and then leave them sitting whereas the ATN will try harder to get the right answer.

V.1.2.2 Marcus' 'Wait-and-See' parser

The 'wait and see' parser (WASP) has similarities both to our surface parser[4] and to Kaplan's ATN. Like our parser, WASP depends on creating small, simple structures and eventually combining them to form the more complex ones, and, like the ATN, it proceeds from the beginning of the sentence to the end without going backwards and forwards in looking for complex constituents. WASP works roughly as follows:

1) The parser proceeds from the first word of the sentence and looks for matching patterns (sometimes predicting what should be the next element in the sentence).

---

[4] Our parser has borrowed somewhat from Marcus'.

2) If the current word does not seem to form part of some constituent (does not match a pattern) it is stacked on a list of constituents.

3) If it finds a matching syntactic pattern, the words/constituents comprising that pattern are phrased together and treated as a single constituent. After a constituent is parsed, it is put on the list of constituents along with unparsed words.

4) Forming a new constituent can, by matching another syntactic pattern, cause already parsed constituents to be popped off the list (last-on first-off) in order to be incorporated into some new pattern.

5) A constituent can have various kinds of information associated with it and its sub-constituents. In a noun phrase, the entire phrase might be labeled[5], and the head noun in that same phrase might be so labeled. In a sentence, the subject and object(s) might be labeled for their function as well as for their syntactic structure.

In a sense the WASP employs a different organization of practically the same basic notions of patterns and parsing style as are used in our surface parser. We can imagine recasting our several passes across constructed constituents into interlocking patterns, where the formation of one constituent allows another pattern to activate itself and try to find all the necessary components. Also there is an appealing sort of psychological plausability to this parser (as well as Ron Kaplan's ATN) in the left-right direction of attention focus movement, and the single pass through the sentence. There is no good evidence to support the opposite view that attention scans back and forth through the words and constituents of the sentence in the way that our parser would have it done.

---

[5] It could receive additional labels as part of further phrasing.

V.1.3 Semantically guided parsers

The parsers we have discussed here are all syntactically based. The patterns upon which recognition depend are independent of considerations of 'meaning' of the words or sentences involved, or of the structure of the discourse up to that point[6]. Interesting work on parsing using 'world knowledge' or reference to meaning has been done, especially by Winograd and Schank[7] who, despite the differences between their approaches, share the underlying assumption that non-structural information is important in structural analysis, that syntax is not self-contained but requires semantics (and possibly additional sources of information) in order to be analyzable.

---

[6] They are also independent of the weather at the time of the parse, but we take it as evident that the reasonableness of 'meaning' in parsing does not require the justification that reference to the weather does.

[7] Samples of different approaches along these lines are contained in Schank and Colby (1973).

We have not closely looked into the use of semantically oriented parsing for two reasons. One reason is simply the practical one that this more complicated parsing requires more resources and has not been implemented in a general enough way that it could be a candidate for incorporation into this project. Such a parser would have required at least a dissertation's worth of work by itself. The other reason for not pursuing semantic parsing is our belief that the phonology (intonation) can utilize only syntactic structure along with some very specific stress/destress information, but without access to semantic information to guide the elaboration of tone groups. If the parse itself depended on semantic information any claim in this area would be a vacuous one.

V.2 The current parser

Our surface syntax has relied on the regularity of English, as expressed by phrase structure systems like X-bar, and in constraints on transformations, to give plausability to our parser. A large segment of the structures utilized in this surface parser were 'discovered' by considering transformational structures and relationships (both deep and surface). Each of our parsing rules tries to resemble a generalization about English syntactic structure. The degree to which we are free of particularities and peculiarities found in English sentences is largely due to the concentration on these generalizations.

One rich avenue of connection between transformational structures and surface syntax is the X-bar notation for phrase structure rules (Jackendoff, 1974). The X-bar notation collapses phrase structure rules for different categories into similar structures as a way of limiting phrase structure systems. Combining this limited "deep structure" with the constraint on cyclic transformational rules that they must be structure preserving, i.e., that they do not change any node labels created in the deep structure, we can see that a good deal of the limited phrase structure survives to be(come) the surface structure of the sentence. Chomsky (1977) combines the X-bar system with a highly limited set of transformations which are subject to strong constraints on their application to structures. These transformations also do not create new structures as a result of application, but only move constituents from node to similar node.

Another feature of this parser is that it always produces one structure for a given string of words. Possible structural ambiguities are ignored in producing a parse. In a sense, this feature reflects the dominance of normative structural patterns built into the parser. A complaint could be made that it is necessary to produce multiple parsings for ambiguous sentences. For example, the phrase "The old men and women" is ambiguous between a reading in which the women are 'old' (in which case the conjunction only joins the nouns, 'men' and 'women') and a reading in which the age of the women is unspecified (the conjunction joins two noun phrases, 'the old men' and 'women'). A parser which produced only one parse of this phrase would need some mechanism for deciding which reading was correct in sentences which contain the phrase. If this mechanism is not present, and we have no concrete suggestions for realizations of this sort of mechanism, a reasonable alternative is to produce alternative parsings of the phrase and allow the procedures that use the structure to make the decision. Such other procedures might have access to some additional information that might clarify the issue. Our defense for producing only one parsing for a given sentence is that, in light of our goal of producing a parsing that will be useful for generating prosodic structure in an actual synthesis system, we try to pick a parse that will result in prosodic assignments that will be ambiguously perceived after synthesis. In the above example, the parse using the noun conjunction will yield the more ambiguous sounding prosodics.

V.2.1 Parser description

In the figures below, we list the patterns that our parser recognizes. In general, the parser uses only the parts of speech of the words in the sentence and not the words themselves. Occasionally, specified words are used in the parse. These are noted in the figures by being quoted. Single upper-case letters stand for specific parts of speech; lower-case letters are part of speech variables. The dash, (-), separates elements of a construction. An exclamation point (!) after a letter or constituent indicates unlimited repetition of that letter or constituent is accepted. To satisfy a pattern, at least one instance of the repeated element must be present. Conditions on the elements in the construction are specified in the area to the right of the colon. The several possible values for variables used in that line are enclosed in angle brackets and separated by a comma. Other conditions on the applicability of a pattern are also listed after variable values. The numbers of the patterns indicate the order in which they are scanned; no pattern can apply after another with a lower number. Where several patterns have the same number, e.g. 2a, 2b, 2c, they are exclusive alternatives.

The parser works by inserting brackets around existing constituents and then assigning a label to the newly created constituent. Each word, at the beginning of the parse procedure, is a separate constituent, whose label is its part of speech, before the parse begins. Several separate scans across the sentence occur during a parse. The parser normally looks at labels starting from the leftmost and proceeding across the string. For some rules, the order of scanning is reversed: the rightmost label is scanned first. When a pattern is matched, the symbol to the left of the arrow (<=) is the label assigned to the resulting bracketed phrase.

{SIMPLE NOUN PHRASE}

1. N <= [N-X-n] : X = <"AND", "OR">
n = <N,R>

2a. N <= z-[A!-N] : z neq D , z possibly null

2b. N <= [D-A!-N] :

2c. N <= [D-N] :

3. P <= [P-n] : n = <N,P>, if n=P then P must dominate a [P-N]

{SIMPLE VERB PHRASE}

4a. V <= [c-v] : c = <C,V>, v = <A,V>

4b. V <= [c-[a-v]!] : c = <C,V>, v,a = <A,V>, a neq v

{HEADLESS NOUN PHRASE}

5a. N <= [A!-A] : scan is reverse

5b. N <= [D-A!-A] : scan is reverse

6. P <= [P-n] : n = <A,P>, if P then P must dominate a [P-A]
scan is reverse

Figure 34. Part 1: Summary of constructions parsed.

The patterns are grouped according to the type of constituent that is involved in the pattern. The first group (1-3) revolve around nouns. Rule 1 forms simple conjuncts: a noun with another noun or pronoun. The case of a pronoun and a noun phrase will be caught later (rule 9). The rules in 2a-c form a noun phrase out of an optional determiner followed optionally by an unlimited number of ad-words and then a noun. A bare noun is not bracketed by these rules. Rule 3 forms a prepositional phrase out of a preposition and noun phrase or a preposition and a prepositional phrase that is not a simple preposition.

Rule 4 forms verb phrases from verb and/or ad-word sequences. 4a is the case of a single verb or copula followed by a verb or ad-word. Pattern 4b creates interior bracketings when the verb or copula is followed by ad-word-and-verb or verb-and-ad-word sequences.

When the simplest noun and verb phrases have been bracketed, the parser looks for adjective phrases that are similar in structure to the simplest noun-phrases (rules 2,3). The parser scans in reverse to give the widest inclusion to these headless noun phrases, phrases missing their head nouns, as in the underlined phrases of these sentences "Plato loved the good", "This is the very best we have to offer," or "Mary wanted the blue one but John preferred a green."

{HIGHER NOUN PHRASE}

7. N <= [N-n!] : n = <N,P>

{CONJUNCTS}

8. E <= [X] : X is sentence initial

9. u <= [u-[X-u]] : u is any phrase, all u's are the same

{HIGHER VERB PHRASE}

10. v <= ["TO"-v] : v = <V,C>, "TO" is a P, scan is reverse

11. V <= [v-p] : v = <V,C>, p = <P,V>, scan is reverse

Figure 35. Part 2: Summary of constructions parsed.

The next several rules create slightly more complex structures using the constituents created (and not created) by the first six rules. Rule 7 forms noun phrases from noun phrases and their complements which can be either other noun phrases or prepositional phrases. These structures produce noun-noun compounding as well. One example of a noun phrase complement to a noun phrase is the double object construction, as in

```
Bill bought   the boy on his block    a blue bicycle.
 N     V [N [N [N    ] [P          ] ]  [N             ] ]
```

In this example, the phrase [the boy] has a (first) complement which is a prepositional phrase, [on his block], and a (second) complement which is the second object of the verb "gave", the noun phrase [a blue bicycle].

Rule 8 separates sentence initial conjunctions and complementizers into their own phrase. The "phrase" constructed at this step does not enter into any other patterns.

Rule 9 forms conjuncts between any two similar phrases, calling the resultant phrase by the same name. An internal bracketing is set up so that the right conjunct and conjunction will be a unit. If there are more than two conjuncts, as in the example below, then they do not all embed at the same level. The first two will be the deepest embedded in the structure[8].

---

[8] An alternative parsing could have a non-binary structure and embed all the conjuncts at the same level.

```
I        came and saw   and conquered.
N      [V [V   [X  V ] ] [X  V       ] ]
```

Rules 10 and 11 form extended verb phrases. Copula-verb sequences get bracketed here, and "to" is considered a copula in this case. Participles are attached to verb sequences by rule 11 (participles are marked as prepositions). If the 'P' phrase is an actual prepositional phrase, then that phrase is an appropriate (first) complement to the verb.

(HIGHEST NOUN PHRASE)

12a. n <= [[X-X][n-[v-n]]] : n = <N,R>, v = <V,C>, scan is reverse
12b. n <= [[X-X][n-v]] : n = <N,R>, v = <V,C>, scan is reverse
12c. n <= [X-[n-[v-n]]] : n = <N,R>, v = <V,C>, scan is reverse
12d. n <= [X-[n-v]] : n = <N,R>, v = <V,C>, scan is reverse
12e. n <= [n-[v-n]] : n = <N,R>, v = <V,C>, scan is reverse
12f. n <= [n-v] : n = <N,R>, v = <V,C>, scan is reverse

(HIGHEST VERB PHRASE)

13a. v <= [v-n!] : v = <V,C>, n = <N,R>, scan is reverse
13b. v <= [v-d!] : v = <V,C>, d = <D,X>, scan is reverse
14. N <= [d!-v] : v = <V,C>, d = <D,X>, scan is reverse

Figure 36. Part 3: Summary of constructions parsed.

The rules in 12 construct complement sentences and independent clauses. If there is a single or double complementizer, it is inserted into the beginning of the clause, which is labeled "N".

Examples of double complementizers are cases where a conjunction and complementizer occur sequentially. This construction is an instance of relying on general structural criteria to provide good parses in cases where a more detailed analysis might have just enough information to make a mistake. Other parts of rule 12 parse clauses, with and without complementizers, thus including full sentences and relative clauses, both full and reduced.

The purpose of the rules in 13 is to form verb phrases that, for whatever reason[9], were not parsed by rule 12. The rationale for looking for determiner/complementizer sequences is that some deletion rules, in transformational terms, will strand these elements as left-over from noun phrases.

A similar rationale applies to pattern 14. Due to multiple categorizations for many words, as well as derivational deletions, a stranded determiner could act as the subject of a clause or sentence. The sequence of determiners could be the result of a determiner to a determiner, i.e. a determiner phrase as in "the boy's book" whose bracketing could be written as [[the boy]'s book]. The possessive indicates that the phrase, [the boy] is a determiner sub-constituent (as in "John's book") while the word "the" is also a determiner. A sentence such as "I gave Bill my book while John got the boy's." shows that the stranding of a pair of determiners is possible.

---

[9] Some reasons are: imperatives are subjectless and therefore not parsed as sentences by 12; gerundive clauses occuring as the object of a verb do not have the structure expected in 12.

V.2.2 Parts of speech

We have claimed that our parser is 'simple' because the structural relationships which it uses are almost all stated in terms of parts of speech (POS) rather than more detailed features of words. One more complicated lexical feature that we have avoided is sub-categorizations. In some current accounts of syntactic structure[10] the lexicon, which can be very important to grammaticality judgements, contains sub-categorization contexts for each word in the lexicon. These sub-categorization are similar to the patterns which we employ in our parser but the difference is that in specifying this information at the lexical level, individual exceptions to the general form of these patterns can be easily included, for example, by listing only those patterns which actually apply to a given word. While there are obvious advantages to using lexical sub-categorizations (easier and more precise corrections to mistaken parses, for example) we have taken the more general approach.

Another possible complication that we have avoided is marking words as exceptions to certain patterns. Marking words is related to the idea of sub-categorizing them, but is conceptually more restricted. In a simple marking scheme, Words would be marked either to never match in a given pattern, to always match in a given pattern (regardless of what other words had to mis-cast), or to sometimes match in a given pattern (the normal, or unmarked case). For example, our punctuation scanning (see below) already uses a version of this feature in marking phrases as "reduced forms" so that they act like nouns except for a few elementary rules. Another way in which we have utilized the idea of exceptions is with the words which appear in the patterns. Using a particular word or small set of words in the pattern is the same as using a more general pattern and marking the word or words to which that pattern applies. So far, we have not felt compelled to implement any further form of exception features.

---

[10] Cf. Chomsky (1965), Wasow (1976).

V.2.2.1 How we define the POS for a word

It could be claimed that we are actually cheating in the way in which we use the POS. If we had a large number of POS', most of which appeared disjunctively in patterns, this claim would be fairly to the point. For example, if we had POS' N1, N2, N3, N4, and, a pattern N1 <= [n V o], where n could be N1, N2 or N3 and o could be any of N2, N3, or N4 we might suspect that an exception feature had been disguised into the creation of four subcategories of noun, which were very similar.

If we carefully adjusted our patterns and assignments of POS so that sentences parsed properly an accusation of "impropriety" in claiming linguistic significance for this parser could be leveled. What, then, are the criteria by which we determine the POS for each word[11]? The basic criterion for assigning a POS is the acceptability of the word in a simple frame (sub-categorization). Nouns are identified by the possibility of taking determiners ('the', 'a' and others). Proper nouns are not subject to this criterion, they are automatically nouns. Reduced forms include all pronouns, and contractions except for some auxiliary-contractions. Determiners are all words which can determine (precede and form a phrase with) nouns. Possessives are all considered determiners. Prepositions are words which can take a determined noun phrase as object. Ad-words (in particular, adjectives) can appear between the determiner and head noun of a noun phrase. Other ad-words modify verbs and entire sentences. Auxiliary verbs are a small closed set containing forms of 'be', 'have' and modals, as well as contractions of these words with each other and with 'not'. Contractions of these words with nouns and pronouns are considered to be reduced forms for the parser. A verb must be able to appear in a simple subject-verb-object(s) pattern. All conjunctions, and complementizers (as well as a good many words that don't really fit into the simple patterns given above) are grouped together. Conjunctions are easily identified by the symmetric constituents that can appear on each side. Complementizers must be able to appear clause initially and not be of any of the other categories described. A proposal for a POS designating wh- words is conceivable, but there does not seem to be a great usefulness for it.

---

[11] We are describing the theoretical criteria. In practice most cases are decided by using some dictionary's assignment of POS.

V.2.3 Punctuation scanner

The parser we have just described makes no use of information about the structure of a sentence which is not present in the words of the sentence. There are many cases where a word is used to name itself or for a phrase to act like a complex noun regardless of its structure. These phrases and words are usually enclosed in punctuation marks which set them off from the rest of the sentence, e.g., colons, double dashes, quotation marks and parentheses. A simple case is the following: "The word 'hit' is a verb." Here the quoted word (hit) is a verb but plays the role of a noun in the syntactic structure. Similarly, the enclosing phrase above (the word 'hit' is a verb) is set off from the rest of the sentence by a colon and quotation marks. Although it is a complete sentence in itself it acts like a name (noun). We have just used parentheses for similar purposes several times, converting words and phrases to act like nouns.

It is important that phrases retain their internal structure even with their changed structural role. We use a punctuation scanner that can recursively call the parser. When the punctuation scanner identifies a sub-phrase it uses the rest of the syntactic analysis to obtain a bracketing for that phrase and then incorporates that bracketed unit into the original sentence as a phrase/word with its part of speech being "reduced noun" so that it can not take adjectives and determiners, but otherwise is treated as a noun.

Punctuation marks like period, question mark, exclamation point and semicolon are all essentially sentence terminators. With a few exceptions these punctuation delineate full syntactic units. Exceptions are that periods are used in abbreviations and the relationship of the terminators and quotative punctuation is somewhat fluid. Periods (and other terminators) can (whether they are supposed to or not[12]) occur inside of quotes or outside of them. Parsing sub-sentence units would be easier if all full-sentence punctuation were outside of the quotes while all strictly sub-sentence punctuation were within, but (alas) that is not the case. Our scanner employs some simple strategies for guessing at the structure of punctuation. For example, if a period does not have spaces to its right it is assumed to be part of an abbreviation. This strategy fails in some cases but is somewhat successful for cases like "i.e., ..." where the first two periods are not treated as terminators. In fact, only the last period is a terminator in that phrase.

Commas are completely ignored in the current scanner. The reason for ignoring them is that they occur in several hard to distinguish contexts[13]. They are used to separate list items, of whatever length, to set off appositives and sentential adverbials, and, together with conjunctions, to separate the final element in a list. One simple possibility for using information from commas is to relate them directly to the phonology by forcing an opposite tone[14] on the comma'ed word and to insert a noticeable pause at the comma. The suggestion derives a casual description of "what to do when you are reading and come to a comma." Neither part of the suggestion has been adopted pending a closer study of actual cases. Another ignored part of the information available in written English is the use of capitalization. We have no suggestions on how particularly to use capitalization other than as a check on where a sentence begins.

---

[12] The American College Dictionary says that periods should always occur before the final quotes (in America but not in England), semicolons always after, other terminators dependent on meaning.

[13] We have not been able to research the differences enough to distinguish cases.

[14] In a contour in which the high tone is prominent the opposite is a low tone, and vice versa.

V.2.4 Unparsed constructions

The parser is basically a local parser in the sense that it never looks more than a few constituents[15] ahead or behind of the current parse focus. The punctuation scanner is not a local procedure in this sense, since in order to match quotation marks it scans as much of the input string as necessary. Another use for a non-local scanner is to identify large scale constituents without having built them up during a previous parse. The earlier suggestion about dividing sentences at their conjunctions is one example of using non-local scanning, but it is fraught with difficulty as pointed out in the earlier discussion.

It might be possible to successfully parse "if-then" sentences by using a non-local scanner. Given a sentence with the words 'if' and 'then' present (and in that order) we could use recursive calls to the parser to achieve a bracketing like:

```
[       ["IF"   [alpha] ]
        ["THEN" [beta] ]        ] .
```

---

[15] Constituents are treated as unitary objects whether they are single words or long phrases.

Alpha and beta are the parsed sub-constituents of the sentence. We can get this non-local scanning by inserting parentheses around the 'if' and 'then' constituents and then letting the normal quotation scanner handle them. We have not currently implemented this idea, because there may be more such constructions which we would want to handle in a similar way (e.g., either-or, so-that, as-as, more-than).

Chapter VI

Semantics and text-view

It should be fairly obvious that access to only the syntactic structure of a piece of text is insufficient both for realistic speech synthesis and for an interesting theoretical treatment of intonation. Examples of this insufficiency are discussed below in detail, but some topics that need discussion are non-neutral stress placement due to local and global semantic influences and speed of delivery, loudness, and pause length as functions of discourse and pragmatic rules. We have, somewhat arbitrarily, selected two additional "components" for our model as playing a significant, definable role in intonation. Our criteria for separating them is that one component will be "allowed" to interact with tonal elaborations inside of sentences (and phrases) while the other will affect the tonal elaboration in such larger units as paragraph, discourse or story scale. The first of these will be labeled "semantics" on the rough similarity to classical semantic concerns. The second scale of effect is labeled "text-view", after Abelson (1973) who introduced it for the analysis of language phenomena of a similar scale. While we achieve a desirable theory in which we have split the intonation assignment responsibilities between these components, we do not (at least at present) intend to claim that our divisions have any special status. An alternative, compatible model of tone group elaborations leading to intonation assignments could probably also be constructed using a single, massive "structural definition" component which could combine what we have labeled "syntax", "semantics" and "text-view". O'Shaughnessy (1976) segregates phenomena differently in his account of intonation, labeling some phenomena to be semantic where we have considered it syntactic, and vice versa. It seems that one person's syntax is another's semantics and perhaps a third's pragmatics, etc. We are simply constructing a theory (and practical system) that will hopefully be adequate for describing much of what is involved in intonation.

VI.1 Semantics

Let us agree from the outset that a complete semantic description of a sentence is not only difficult to achieve but includes many elements which have no bearing on intonation synthesis. What we are concerned with are instances where some non-syntactic (this being our definition of "semantic" for the purposes of this discussion) phenomena influence elaboration of the tone group and thereby the prosodic assignments. While it is probably true that no two sentences ever have exactly the same meaning (semantic content), we will be concerned with differences that affect the elaboration of a tonal pattern[1]. The tone group elaboration can be affected either locally or globally. Local changes to tonal elaboration are usually perceived as simple cases of stress. Global changes, while also having a component of perceived stress, can be said to cause a change in focus and overall intonation contour.

---

[1] It is our belief that at least some of the variation in observed tone patterns derives from small semantic differences. These differences can involve personal meanings inaccessible to others as well fleeting realizations of meanings. We will not delve further in this direction, but will treat such differences as random variation.

VI.1.1 Local stress and de-stress

We discussed above (in Section II.5) what we consider to be the nature of emphatic stress in English and (in Section III.3.2.1) what it might mean for a word to be destressed. Our tone group hypothesis makes the stressing of both single words and entire phrases seem similar. At some level of elaboration/analysis phrase stress should look like single word stress[2], and, within a stressed phrase neutral tone group elaborations should take place, relative to the stress.

Semantic phenomena of this type can be classed as primarily lexical, primarily contextual or mixed between the two. In the more lexical cases the meaning or semantic 'effect' of a particular word causes a perturbation in the neutral intonation assignment in that word's portion of the pitch contour. The words 'always' and 'never' exhibit what we consider to be lexically determined (semantic) emphatic stress[3]. In Figure 37 we show some sentences in which these

---

[2] Our current synthesis system does not include this generalization.

[3] We are being somewhat arbitrary in separating syntax and semantics at this point. We chose to accept as syntactic only structures that are extremely general, choosing to complicate our semantics with cases that have less generality. The choice could have been otherwise and we do not see any major purpose in fighting over the definition.

1.	I		ate	a	few	of	the	cookies	
	93		112	111	122	100	-	112	al
	95		114	84	96	86	83	110	gb
	238		244	238	250	238	233	244	do

2.	I	never	ate	a	few	of	the	cookies	
	100	137	102	-	115	105	93	106	al
	91	115	101	89	99	106	85	99	gb
	244	303	263	222	263	238	222	238	do

3.	I	always	eat	a	few	of	the	cookies	
	106	143	108	104	119	100	92	115	al
	96	105	98	99	103	94	84	102	gb
	244	294	263	213	270	233	204	233	do

Figure 37. Some observed pitch contours.

words produce a stress contour. Each of these sentences was spoken by three different speakers (al and gb are males, do is female). The sentences given here were part of a set of twelve sentences; these particular sentences were not recorded one after another. In sentence 1 there is an upstep from 'I' to 'ate' and an elaboration of the partitive that fits the syntactic structure ([[a few] [of [the cookies]]]). Sentences 2 and 3 show a similar pitch elaboration in the partitive phr[illegible]e. The words 'never' and 'always' show a local elaboration of a high tone as we described above (Section II.5), and, the emphasized word has an isolated pitch peak. The phonetic differences (never/always, eat/ate) do not seem to be an important factor in the assignment of pitches, though the word 'eat' seems to be pronounced with a slightly higher pitch with respect to the following word than 'ate'.

An example of a largely contextual phenomenon is the use of quotes around a word (or phrase). The quotes are a context in which the quoted word usually receives a stress contour, regardless of its lexical function. One explanation for giving stress to quoted material is the extreme freedom of substitution within quotes. While syntactic and semantic constraints make phrases somewhat predictable in neutral utterances, few of these constraints apply to quoted material. Following this explanation further, we expect that in general, in cases of semantic and syntactic unpredictability a stress contour will be required. Some other contexts that normally exhibit this sort of unpredictability are cleft and pseudo-cleft sentences, such as "What I want is x". These constructions have an effect of making the material (labeled x here) after the 'introductory' part of the sentence (the "what I want is") fairly unrestricted (and thus unpredictable) as to internal syntactic structure and semantic content.

We saw an example of (what could be called) a mixed case of lexical and contextual stress above (Section III.3.2.1), where the verb 'requires' exhibited the upstep that we attributed to its being de-stressed to the status of an auxiliary, semantically equivalent to 'be' or 'must'. The context provided the condition for a semantic reinterpretation of the lexical meaning and prosodic contour. Our notion of context is also tied up with focus which we will discuss below.

VI.1.2 Global contour changing

In the introduction, Section I.1.2.1, we mentioned studies by Liberman and Sag (1974), and Bolinger (1972). Their work points out semantic concerns that do not simply result in stress (or, tone group elaborations producing stress). Liberman (1975) also discusses a number of different intonation contours and connects them to semantic interpretations. We will give one example of a semantic interaction with prosodic elaboration. We did not discover this example, but only report on it and translate it into our own terms.

VI.1.2.1 focus shift

A great deal has been said (and written) about topic/comment (semantic) structure of sentences[4]. For this discussion, we will skirt the difficulties of topic/comment and take a simple (intuitive) definition of the "focus" of a sentence as the constituent which relates most closely to what is important in the sentence. The use of "most closely" and "important" in the definition threaten to be a can of worms if we try to specify them too closely. We will not pursue the matter that far; all that we will use here is a vague acceptance of the related idea of "shifted focus". In different utterances of a sentence, and in utterances of similar but different strings of words, what is most important in the utterance can change. Usually the focus (the most important constituent) of an English sentence is its subject

---

[4] While focus -- topic/comment -- is normally classed as a discourse phenomenon (and thus not strictly semantic), under our definition it is still semantic because it affects the elaboration of the tone group inside of a sentence.

(or a preposed constituent as in sentences with topicalization), but, depending on the discourse context or the speaker's intention, a constituent occurring elsewhere in the sentence can be focused on. There is a connection between focus and stress in a sentence, but the expression of focus can be separated from the production of stress and linked to the general elaboration of tone groups.

Let us introduce the notion of "prominence". Goldsmith (1974, 1975), Liberman (1975) and others elaborate their tones with the prominent element (word/constituent/syllable) of the utterance matched to one tone of the group which is marked as "prominent". In the tone group "declarative fall", (M)HL, the high tone is marked prominent. For a sentence in which the subject is the focus (which we will say is the prominent element) the high tone is associated with the subject. Our theory of tonal elaboration assumes that subject position is a normal focus position in a sentence[5]. When the focus is shifted, an association of the focused constituent with the prominent tone (high tone in a declarative) is a necessary part of the tonal elaboration. In the declarative contour, this will usually make the mid-tone extend over more than just the single word that would normally be within its scope to cover entire constituents. We diagram the situation below, Figure 38.

Another contour shifting effect can be linked to the focus. Constituents which closely precede the focus (especially in a case of shifted focus) have their relative semantic weights equalized. For example, the relative weights of a noun and its determiner (which normally has a much smaller weight) can be evened out. This relative (semantic) importance of the determiner is reflected in the non-realization of the optional mid tone as part of the elaboration of the sub-contour. Often some determiner (or auxiliary) in a constituent immediately preceding the focus constituent will not upstep to the head of its phrase, as would be expected in a neutral elaboration. In general, when an entire phrase consists of auxiliaries, for example when the main verb of a conjoined clause is deleted, then the first auxiliary does not upstep to the last one. The situation in a noun phrase preceding a focused constituent is similar to an auxiliary sequence since the head of the phrase has been somewhat de-stressed due to the focus shift.

---

[5] Thus the predictions of intonation contour from a Liberman/Goldsmith elaboration come out similar to our syntactically guided elaborations.

Figure 38 shows some of the effects we have discussed in our abstract tone notation. We designate the focus constituent with a star, '*' (following Goldsmith (1974,5)), over the high tone which we associate with the focus. In the first version of the sentence the prominence is associated with the subject ("the man"). In that case we get a neutral tone elaboration. At the highest level of that elaboration we simply get a downstepping from the high tone on the noun phrase through the other constituents of the sentence[6]. The elaboration of the initial high tone produces a mid tone on 'the'

---

[6] This example does not crucially depend on whether we consider "crushed the old cars" to be a single phrase or two separate phrases.

1.

```
   *
The man crushed the old cars.
   H     H-1        H-2      L
 M   H L          M   H'  H'-1L
```

2.

```
           *
The man crushed the old cars.
   H       H        H-1      L
 H' H''L          M   H'  H'-1L
```

Figure 38. Some schematically shown effects of focus shift.

because it is a minor lexical item. In the second version of this sentence we have indicated a shifted focus. This shift might result from uttering this sentence in response to the question, "What did John Doe do with the cars?" By shifting the focus to the verb, 'crushed', we alter the elaboration of the tone groups, both the one for the entire sentence and the one for the initial noun phrase. For the entire sentence we have forced a mid-tone at the beginning of the sentence, and, within "the man" the expected upstep is not realized because the relative semantic weights of the determiner and head have been equalized[7].

---

[7] We might even want to claim that in answer to the question, "What did John Doe do with the cars?", the subject noun phrase (and, in particular, the noun 'man') is acting like a pro-form. Talking about a "pro-form" makes the focus shift a reasonable idea, and helps explain both why the entire phrase is assigned a mid tone and why the determiner does not upstep to the head (pro-form) noun.

1.	I	ate	a	few	of	the	cookies
	238	244	238	250	238	233	244
2.	I	ate	very	few	of	the	cookies
	263	250	312	294	238	222	250

Figure 39. An example of shifted focus.

Figure 39 gives an instance of the effects of shifted focus. Both sentences were recorded by do (a female reader) among a list of other similar sentences (other examples were given above, Figure 37). The difference between the sentences is the substitution of 'very' for 'a'. Without a notion of focus shift, we would have expected only the different elaboration of that single word. What we observe is a double effect. First of all, the peak pitch of the sentence moves from 'ate' in sentence 1 to 'very' in sentence 2. One facet of this shift is that 'very' gets a higher than normal[8] pitch peak. Notice that, while 'few' has a higher peak pitch in sentence 2 than in sentence 1, the phrase "of the cookies" has pitches that are similar in both. This whole aspect of the focus shift closely resembles emphatic stress -- it is local and affects the pitch assignments similarly. The second aspect of focus shift is the change in relative pitch of the first two words, "I ate". In the first sentence 'I' upsteps to 'ate' whereas in the second it downsteps. By focusing on 'very' we have evened the relative semantic weights of 'I' and 'ate' so that we no longer have the context for an upstep. We described a similar evening out of semantic weights before a focused constituent above in talking about the sentence "The man crushed the old cars". This is the difference between simple stress and a more global semantic change as exemplified in focus shift.[9] Bolinger (personal communication) has pointed out similarly that in a sentence like "John typed the answer", the focus of the sentence can fall on any of the three constituents. If the focus was on "the answer" we would expect that the usual upstepping of 'the' to 'answer' could be changed to a downstep because of the increased semantic weight on 'the'[10].

---

[8] In the full list of 12 sentences recorded by do the normal high pitch for a sentence is about 260-280 herz.

VI.1.3 Neutral and non-neutral contours

We have just seen some instances of what might be considered "radical disruptions" of the neutral contour which we developed in our description of tone group elaboration. In the lexical stress cases, the normal (most often occuring) intonation contour is not our syntactically based tonal elaboration, but rather something that we described as a disruption of that elaboration. The global

---

[9] It is possible to construe the evening out of semantic weights which we are using as being a facet of the rhythm of a sentence with a (shifted) focused element. On this view, focusing on 'very' brings with it a rhythmic structure to the sentence in which the words 'I' and 'ate' are said off the main pulses of the sentence. Because they are both "off the beat", and not because their semantic weights have been changed, 'I' does not upstep to 'ate'. Our view is that the rhythmic structure of this sentence is also a function of semantic weight (e.g. function words are shorter and not on the main pulses of the sentence, in general, because they do not have semantic content) rather than being independent and simply parallel to it.

[10] This downstep is in fact observed, see above Section III.1, sentence "Please type the answer."

"disruptions" were even more severe in altering the pattern of elaboration. Does the possibility of radical disruption of the normal elaboration of a tone group due to semantic interaction negate the value of having a neutral contour? Schmerling (1973) argues against a characterization of "neutral" stress contours being possible without some circularity or over-simplification. Either a neutral contour is defined to be the output of "neutral stress" rules (circular), or citation form is taken to be "neutral" (over-simplified). Yet she is forced to undermine her own stand.

> It remains true that some utterance-tokens involve more special assumptions than others, and linguists can profitably study the effects different assumptions have on stress, just as we can study what effects various syntactic and phonological properties of utterances have on stress. (Schmerling 1973, p.71)

Moreover, she proceeds (in her subsequent chapter) to define and use a version of neutral contour under the guise of "stressability". Our defense of the "neutral contour" goes as follows:

1) There is a broad class of utterances where the syntactic structure accounts for the tonal elaboration. These are paradigm cases of neutral contours.

2) Even in sentences where a non-neutral contour is the most natural sounding one or where it is the one we observe most often, we can usually identify a non-syntactic (semantic in our usage) source for the

difference. Also, it is almost always the case that even in these sentences, the "rest" of the sentence shows the neutral contour.

3) As a research strategy, a provisional definition of neutral stress as a baseline against which to compare different phenomena and expectations is useful. The base form should be one from which we can derive other observed variants. A major thrust of this dissertation has been to demonstrate some of the potential for a theory of tonal elaboration which is mediated by syntax in neutral cases and by non-syntactic constructs in other cases.

VI.1.4 Why not truth-functional semantics?

The various examples of stress phenomena we have given above are not amenable to a truth functional description. There is no clear way that knowing the truth value of a sentence (or even its logical relationship to other sentences)[11] tells us anything interesting about the intonation contour. Certainly, in the simple examples given in this chapter, we could describe the adverbs and contexts as 'operators' in some logical formalism. The point is that using such a formalism would be a superfluous way of conveying information about the semantic content we are interested in. Even in the case of Liberman and Sag's 'contradiction' contour, the simple contradiction of true/false is not at issue, but rather an intention on the speaker's part to convey his reaction to a preceding question, a non-logical issue.

---

[11] Let us assume that some sense can be made out of the "truth value" of a question or exclamation and similar constructs.

What we really want is a way of measuring and marking information content that would allow us to predict where stress needs to be placed (and where it needs to be reduced) and when a contour needs to modify its neutral elaboration. Our intention here is to provide a notation for semantic structures similar to the representation of constituent structure by the use of bracketing together with parts of speech[12]. The lexical stress bearers that we discussed above can be marked in the lexicon to carry some degree of stress[13]. This marking could be carried over into the representation. Other markings will include a prominence mark (like the star mentioned above) and (relative) de-stressing and re-stressing information. We will also want to include general contour and sub-contour selectors, to provide for different over-all contours like declarative, question or contradiction, as well as sub-contours like the comma intonation.

We might represent this information as part of (really an extension of) the bracketed string used for syntactic representation. For example, suppose that '.', '?', and '!' selected declarative, question and contradiction contours, respectively; that '*' was the focus-prominence marker, and '$' indicated a stress level. Then we could represent some of our preceding sentences as shown below in Figure 40. Sentence 1 is a representation of a sentence with a normal focus ('ate' is the first "content word" in the sentence) and a single stressed word, 'few', in a declarative contour. Sentence 2 indicates a contradiction contour with the subject, 'I', in focus position. 'Few' is still somewhat stressed. In sentence 3 we show the focus shift to the stressed word in the sentence, 'few', while in sentence 4 'very' is focus and stressed while 'few' is simply stressed. The destressing of "I ate" might be explicitly marked in the structure as well. Sentences 5 to 7 show different focuses which correspond to the cases we mentioned above for them. In 5 we have a neutral focus on 'man'[14]. In 6 the focus has moved to 'crushed', but without additional stress. In this case we would expect that the high tone for the sentence would be on 'crushed' but that it would be a normal high tone and not a 'higher-than-high' tone as we would expect in 7, where the word is stressed as well as focus.

---

[12] The problem of parsing a sentence with respect to some system of semantic markings is separate and even less resolved.

[13] A more complicated proposal might have these words marked to trigger a routine which determines whether and how much stress these words will receive.

More work is needed to determine whether this format is useful for representing the information we want. It is possible that a format that makes all of the semantic information available at the beginning of the sentence, as in a separate list which could be ordered differently from the left-to-right ordering of the words. We currently allow a marking of the stress to be placed on a word, more or less in the format indicated above.

---

[14] If we had wished to focus on the phrase "the man" we could have used 5'. [* [The man] crushed [the old cars]], putting the star before the entire constituent.

1. [I * ate [[a $ few] [of [the cookies]]]].

2. [* I ate [[a $ few] [of [the cookies]]]]!

3. [I ate [[a * $ few] [of [the cookies]]]].

4. [I ate [[* $ very $ few] [of [the cookies]]]].

5. [[The * man] crushed [the old cars]].

6. [[The man] * crushed [the old cars]].

7. [[The man] * $ crushed [the old cars]].

Figure 40. Some sentences with syntactic bracketing and amplified for semantic information.

VI.2 Text-view

When dealing with the analysis of sentence structure, we will call upon the familiar concepts of constituency and paradigmatic co-occurrence. These notions have been developed in linguistics to play a central role in the systematic study of forms of language. Constituency expresses the idea that a sequence of words and/or combinations of words can "do the same thing" even though the sequence appears in different contexts. The most common use of constituency comes in discussing syntactic structure and we have used the concept extensively above. Paradigmatic co-occurrence is the related idea that some constituents are substitutable for each other while others are not[15].

---

[15] Syntagmatic co-occurrence is the partner of paradigmatic co-occurrence in that it deals with constituents which can appear with

The use of constituent analysis being proposed here is not an attempt at full understanding of text. That problem is much more difficult than the one attacked here. For understanding a text, much information relevant to real world activity must be included in the analysis. Our goal in this respect is to account for as much of the purely linguistic structure as possible without going to the complication of including real world knowledge. We will try to link the structural relations of "text constituents" to prosodic phenomena. Even so, our treatment of this topic will be somewhat sketchy and impressionistic. In the examples below we will not try to justify our constituent analyses by using co-occurrence facts, although we expect this to be the appropriate methodology for a more completely developed system.

In Appendix III we give the text of a logic lesson, which we have chosen to analyze, and, in Figure 41 we show our (text-view) structural analysis of several of the exercises from that lesson. We treat each exercise (numbered 1 through 11) as a paragraph although some of them probably are more than one paragraph long. The letters used in this figure represent the sentences or sentence groups that the paragraph is divided into. We have given an unlabeled bracketing to these constituents on the (null) hypothesis that they are not required. Since our work on using any bracketing for generating prosodic features is still rudimentary, we can not discount the

---

each other. We will be more concerned with paradigmatic co-occurrence.

1.
[A B C] [[D E] [F G H I]]

2.
[A B C D E]
[F [G H] [I [J K L] M] N O]
[P Q R] S T]

3.
[A [B C] D] [E F] [G [H I] [J K]] L

4.
[[A B] [C [D E]] F] G

5.
[[A B] [C D E F]]

6.
A [B C D] [E F G] [H I [J K L]]

7.
[A B] [C D E F]

8.
[A B] [C D [E F] G] [[H I] [J K]] L

9.
[A B] [C D E F]

10.
A [B C D E] F

11.
[A B C] [D E F] [G [H I] [J K] L]
[M [N O]] [P [Q R]] S T

Figure 41. Constituent structure of several paragraphs

possibility that they might be useful. One important simplification that we obtain by avoiding labels is the prediction that similar prosodic phenomena are found in constituents with different labels. We discussed a similar prediction in regard to our syntactic constituent bracketing where we found that, in general, the constituent labels were not important (Section I.4.2). Our text-view bracketing is not total, unlike our syntactic bracketing. There are some sentences in the paragraphs that are not included in any bracketing. At this stage of investigation we do not have particular reasons for preferring one type to another for the text-view. There are complex constituents in these exercises, where one sequence of sentences forms a unified sub-part of another.

We have chosen a logic lesson for our example partly because of the clarity of the structure of the lessons and partly because the lessons are available and the claims we make about phonological implications can be checked using the recordings that have been made for those lessons at Stanford. We hope that the structures discussed below will be applicable to generalized text analysis, but we must admit that it would be fairly surprising, since there is a certain style to the construction and reading of these lessons (in particular and any (read) lessons, more generally). If further research indicates that interesting continuations of our current ideas are worthwhile we can be quite content. Test questions and other, non-discursive material have been excluded from our examination since they may prove somewhat different from the purely discursive type of texts we are considering here.

Above, in the discussion of syntactic structure and parsing, we said that a large part of the information needed for tonal elaboration (with respect to the syntactic structure) is present in a surface (immediate) structure. Our attempt at text analysis is similarly oriented. If we can isolate structures that are useful for assigning some prosodic parameters we will consider ourselves successful. In particular we have had some success at associating constituent structure with overall shortening or lengthening of the duration of sentences. We will discuss this further below. We have not tried to devise an algorithm for parsing a text into these constituents (or even one for using these constituents as part of a general prosodic assignment system).

VI.2.1 General structure

The basic impetus for the organization we have adopted is a notion of "paragraph structure". We have impressionistically structured the sample lessons into constituents corresponding roughly to: introduction, body, examples, conclusion and footnote. It is clear that any of the assignments which we make can be questioned, much the way one can question constituent structure decisions in syntax. Some degree of circularity, of assigning categories so that the analysis works, is probably present, but we have tried to be fairly honest.

Although we have not tried to establish firm connections between text structure and prosodic features, we believe that some regularities exist. We will describe shortly an observed relationship in utterance speed. Another interaction could be in the length of pauses between sentences. We might expect that pauses between more major sub-divisions will be longer, in general, than those between lower level constituents. An exception to this generalization could be between statement and contrast, where the juxtaposition might be reason for increased pause time. Other prosodic interactions could be found in the pitch interval (step). Again, some sections of text (perhaps the introduction and conclusion, as well as some part of the body) might require a larger step size for their tonal elaborations than others.

VI.2.2 Utterance length

It is fairly intuitive that key sentences (for example, topic sentences) and phrases are said more slowly than the rest of a text. Since determining which sentences are "key" in a paragraph is a tricky task, we will discuss a different regularity that we have seen which links duration of utterance with position in the structural hierarchy. Figure 42 repeats the bracketing of logic lessons (from Appendix III) given above in Figure 41. This time, the numbers displayed inside the bracketings are the difference between the length of the sentence (or sentence group) as predicted by our current duration theory and the observed length of the recorded utterance used in the logic course, expressed as a percentage of the theoretical prediction. A positive value indicates that the theoretical prediction was larger than the observed, while a negative means that the theory predicted

1.
[-11 -16 21] [[-29 6] [7 8 0 -6]]

2.
[-6 -3 16 -5 8]

[11 [11 10] [16 [31 32 32] 11] -10 12]

[4 6 5] 11 6]

3.
[4 [3 17] 13] [-5 -17] [17 [-27 5] [7 7]] 0

4.
[[10 5] [5 [3 4]] -7] 7

5.
[[12 13] [3 3 -4 1]]

6.
3 [12 9 -1] [4 7 19] [11 0 [10 1 3]]

7.
[-4 17] [6 12 1 9]

8.
[16 12] [12 40 [15 14] 11] [[-2 8] [9 6]] 4

9.
[9 18] [7 12 1 4]

10.
3 [-1 6 -5 -4] 15

11.
[18 0 9] [18 8 0] [6 [0 -20] [2 -14] 9]

[15 [30 20]] [-1 [-8 1]] 22 20

Figure 42. Text-view structure: percentage difference of observed length of utterance from length prediction using only syntax and lexicon.

too short a duration for the utterance. If we were trying to model the observed lengths accurately, we would need to shorten sentences which had a positive value and lengthen sentences with a negative one. The method used for these comparisons is similar to that described above Section IV.1.1, in describing the experiment where we tested the overall goodness of our utterance length predictions.

VI.2.2.1 Lists

Clear instances of the regularity we will discuss are in paragraphs 7 and 9. Looking at the corresponding values of these paragraphs we see the similarities easily. The texts of the two paragraphs are also quite similar, both give a two element introduction to a list of four possibilities. We can summarize the observations:

1) The introduction to the list shows a length contrast between the two elements in which the second must be shortened while the first is either lengthened slightly or shortened less. We hypothesize that the first should be predicted close to normal speed while the second is predicted to be slower than normal.

2) In the list itself, the second element stands out as being the most prominently divergent from its prediction. It is predicted to be much longer than observed. We hypothesize that in general (sentential) list elements are said at close normal speed except

for the second element which should be faster than normal.

3) The introduction seems to be on the whole slightly faster than the list, and the list speed is predicted approximately correctly.

Paragraph 5, does not particularly confirm the observations of 7 and 9, though the introduction is predicted to be longer than the observed by more than the list. This may indicate that the whole lesson was read at a slightly faster pace than predicted.

The last major constituent of paragraph 1 shows the contrast within the introduction which we are looking for, and the second element on the list is spoken faster than the rest of the list except the first element. This difference from our hypothesis about list sentence-lengths might be due to causes unrelated to the text-view structure or to aspects of that structure which we have not isolated as yet or may be counter evidence.

Paragraph 10 has only a single element introduction and the third prediction relating the introduction and the list seems to hold. The list structure also seems to follow our hypothesis by showing the second element needing to be shortened while the other elements need to be lengthened from the neutral prediction.

VI.2.2.2 Footnotes

The last value for paragraph 10 is a footnote and is the sentence needing most to be shortened. There are other examples of footnotes which also share this characteristic of requiring shortening: the last two sentences of paragraph 11; paragraph 4; perhaps paragraph 8; not in paragraph 3. Notice that other final sentences do not show the same need for shortening, so we would not attribute this phenomenon simply to being paragraph final.

VI.2.3 Parsing

We have not developed an algorithm for parsing the structures discussed above. Parsing text for these constituents is more difficult than parsing sentences for syntactic constituency because in this case there is no prior fixed assignment of sentences to lowest level categories. We can not say, offhand, whether a sentence is in some body, example or conclusion. One possibility is to use a dictionary of keywords that will enable a rough guess to be made as to the lowest constituency of sentences and phrases. These keywords might also help delineate the scope of the highest level constituents. We can notice that examples often use the word 'example' somewhere in the constituent, usually at the beginning of a sentence introducing the constituent. Using this sort of keywording might allow a fair amount of useful bracketing to go on.

Chapter VII

Technical details of a full speech synthesis system

We would like to describe, in some detail, the actual prosodic processes that we employ in our current intonation generation programs. The easiest way to describe the different parts of these programs is to embed their description into a description of the system that uses the most general form of these programs.

VII.1 TEXTSP, a prototype text-to-speech system

TEXTSP is one of several minisystems produced at the Institute which utilize the current prosodic generation capability. While the original impetus for the creation of these programs derived from the use of limited prosodic generation in conjunction with computer-assisted-instruction at the Institute, the routines needed for that application were found to be suitable for more general use in an experimental context.

VII.1.1 Outline of the program

The TEXTSP system consists of the components shown in Figure 43 below. The text source is a computer-printable file (e.g., suitable for listing on a standard printer or terminal). This file is accessed using the PDP-10 (TENEX) file-handling procedures. As the file is read by the system, one line at'a time, the punctuation scanner removes extraneous punctuation marks, translates other

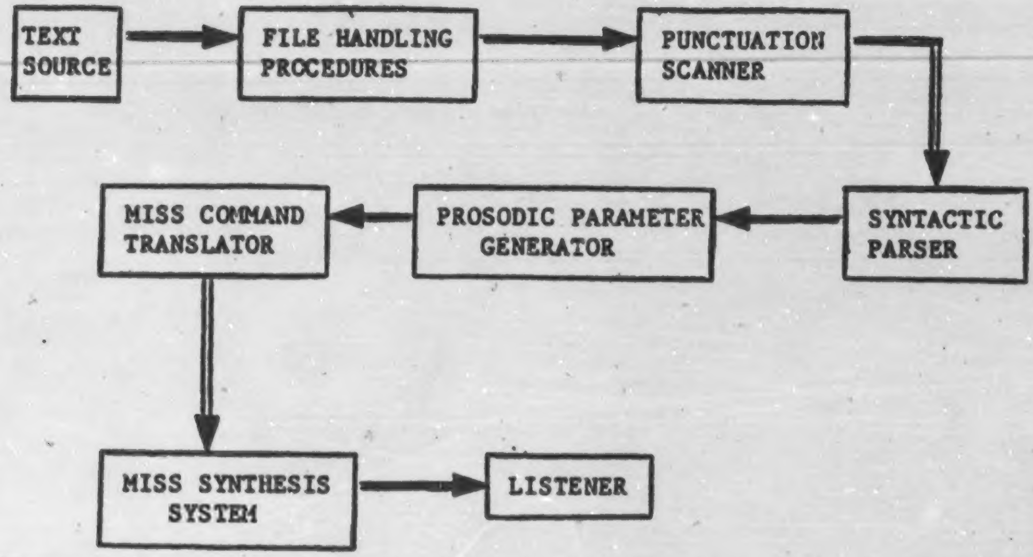

TEXT SOURCE
FILE HANDLING PROCEDURES
PUNCTUATION SCANNER
SYNTACTIC PARSER
PROSODIC PARAMETER GENERATOR
MISS COMMAND TRANSLATOR
MISS SYNTHESIS SYSTEM
LISTENER

Figure 43. The components of the TEXTSP system.

punctuation into lexical items, and structures the input text on the basis of some of the other punctuation. The punctuation component recognizes the standard sentence-ending punctuation and then passes the sentence to the syntactic parser which performs a constituent structure analysis to produce a bracketed string which includes the original sentence words, their parts of speech, and the syntactic structure of the sentence. The prosodic parameter generator uses this bracketed string to decide on the actual parameter specifications for pitch, duration, and loudness, and produces an array of values which are transformed by the MISS command translator into a command string consisting of one line of specifications for each word in the original sentence. In addition, the translator makes whatever compression of the command string is possible, by examining the lines pairwise for repeated specifications. The MISS synthesis system uses the command string to retrieve words from the lexicon and perform the necessary real-time manipulations to produce audible speech. The MISS synthesizer operates asynchronously from the rest of the system; that is, it performs its retrieval and prosodic manipulations of the phonetic data independently of the rest of the TEXTSP system.

Figure 44 below gives a flow outline of processing in the TEXTSP system. Steps 1 through 3 are the user's control over the system. A user can select any ASCII text file to be read; can adjust the speed of reading (typical rates used range from twice normal speed to just slightly slower than normal[1]); and can select any set of pages from the selected text to be read in any order. Steps 4 through 8 give the actual sequence of routines used in the analysis and prosodic command generation phases of the system. As each line of text is read, it is edited for punctuation and syntactically analyzed, and the appropriate parameters are generated into a command string. The MISS machine executes the command string. Concurrent with the execution of the command string by the MISS machine, the system returns via the loop instruction (step 10 in the figure). When the source text file has been read, the program loops (following instruction 11 in the figure) to the beginning to allow selection of additional text material.

---

[1] Cf. Sanders and Laddaga, 1976.

A.

1. Specify text file to be read.
2. Specify reading speed.
3. Specify section of text file to be read.

B.

4. Retrieve desired part of file, one line at a time. Read page numbers as new pages of the file are encountered.
5. Scan for punctuation: Remove or restructure as necessary. Accumulate words to end of sentence.
6. Syntactic analysis of sentence. Includes dictionary lookup of part of speech for each word.
7. Prosodic parameters generated from syntactic analysis
8. Generate command string for hardware MISS sound synthesizer. Command includes word name and compressed parameters.
9. MISS machine executes command string by retrieving words and performing prosodic manipulations.

C.

10. Repeat to step 4, until no more file lines are left.
11. Repeat from step 1, until there are no more requests.

(NOTE: Step 9 is asynchronous from steps 4-8.)

Figure 44. Flowchart/outline of the TEXTSP system.

VII.1.2 How good is TEXTSP?

TEXTSP has been very useful in shaping our conception of the requirements of a general text-to-speech system. It possesses some of the central facilities necessary for such a system, although some of the capabilities are in a somewhat rudimentary form and some components necessary for a full system are absent. An essential feature of the current system is that it always produces an "answer," that is, the algorithms employed yield a speakable string using a limited amount of processing time. Furthermore, the overall system is designed to be lenient in case one component makes a mistake. For example, if a prepositional phrase is ambiguous in a sentence, the syntactic parser will produce the structure that results in a synthesized sound that can be interpreted somewhat ambiguously.

We have noticed some faults in the system. There are cases where the parser produces an incorrect structure. This usually results in an audible error; either the emphasis will sound misplaced or the constituent structure will be missed. Our punctuation analysis is still largely at the ad hoc stage. Many punctuation marks are either ignored or treated superficially. Words that are not found in the lexicon are spelled out, one letter at a time, rather than sounded out. Finally, no semantic and discourse rules or heuristics exist in the current version of TEXTSP.

We expect to complicate our model of structural analysis to include semantic analysis as well as limited-scope discourse analysis as discussed above, Chapter VI, to provide a more complete structure for the prosodic generator to use. In general, our model of a text-to-speech system (Figure 45) will closely resemble the model already in use for TEXTSP.

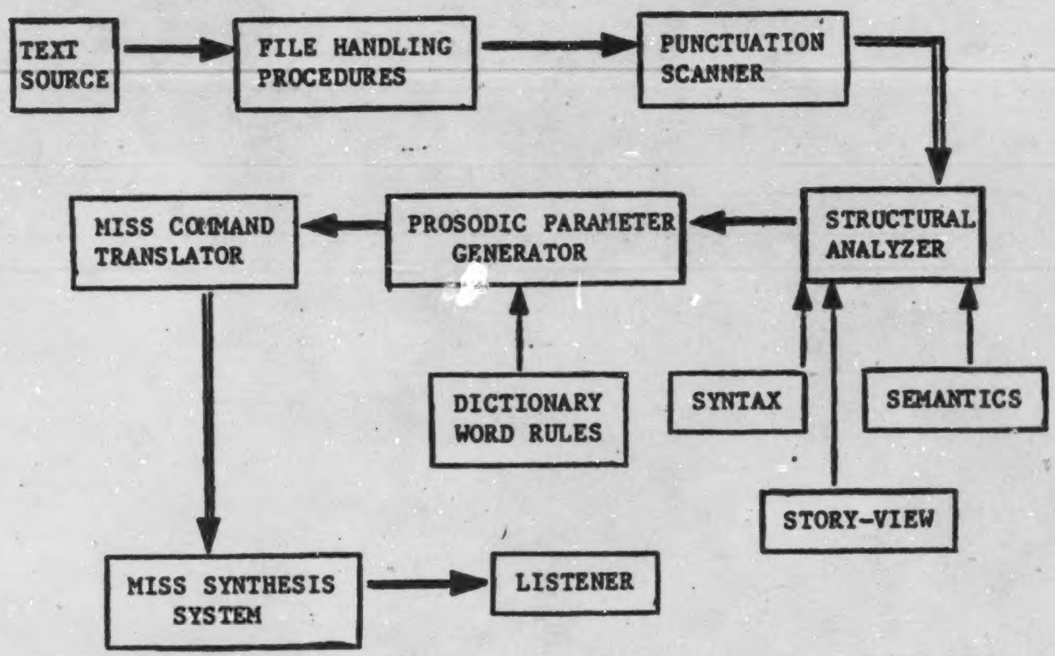

TEXT SOURCE
FILE HANDLING PROCEDURES
PUNCTUATION SCANNER
STRUCTURAL ANALYZER
PROSODIC PARAMETER GENERATOR
MISS COMMAND TRANSLATOR
DICTIONARY WORD RULES
SYNTAX
SEMANTICS
STORY-VIEW
MISS SYNTHESIS SYSTEM
LISTENER

Figure 45. Component diagram for a revised text-to-speech system.

This diagram shows the modifications of the TEXTSP programs to be incorporated in a future text-to-speech system. The important differences between the two systems are that in the proposed system a more elaborate structural analyzer replaces the simple syntactic analysis from TEXTSP. The structural analyzer has as subcomponents a syntactic analysis developed from the one in TEXTSP and, additionally, semantic and text-view (discourse) components. Another important difference between TEXTSP and our proposed system is the use of dictionary word-formation rules. Word formation rules would allow morphophonemic analysis to be done on words that are not stored in the dictionary instead of spelling them out[2]. Also, in this model of the system, the user, either local to the installation or with some remote (e.g., touch-tone telephone) connection[3], can request the reading of some file, available to the file-handling system. With the future (independent) development of optical scanning devices, it would be fairly uncomplicated to add such a unit as an additional means of making (printed) text available to the synthesis system.

Below, we will describe the actual operation of the key components of the system in detail. The theoretical basis for these components was discussed under the appropriate heading above. The syntactic analysis was discussed in enough detail above that we will not repeat the discussion here at all, except to describe the output from the parser.

VII.2 The punctuation scanner

There are three classes of punctuation that play different roles in the current speech synthesis algorithms. The first class consists of punctuation specifically intended to control the operation of the various synthesis routines. This class consists of the non-alphabetic characters shown below in Figure 46 and Figure 47. These characters must be passed to the prosodic generator in order for the appropriate contours to be generated.

---

[2] Cf. Hunnicutt (1976).

[3] An actual computer terminal would not be required for requesting the reading of a file.

The second class of punctuation includes 'period' and 'question mark' and other sentence-enders. It also includes characters like apostrophe, quotation mark, parentheses, and (eventually) comma. The parentheses are used by the parser as a strong form of comma, although they do not appear in its final output. Since they are part of the standard orthography of English, these punctuation are not specified (by a user, when creating a text source) in order to control the intonation of the sentence per se, but they are clues to the syntactic structure. These punctuation marks must be translated into information that the parser and, later the other prosodic routines, can use. For example, a word contained in single quotes (apostrophes) receives a non-neutral stress contour.

Class three punctuation consists of all the possible punctuation marks that have no implications for prosodic generation. Currently, page boundaries, line endings, tabs and spacing are punctuation of this nature. Additionally, for text that was not originally prepared with an eye to speech synthesis, punctuation marks from the first class must be translated into their names, e.g. 'star' for '*', but otherwise play no role in the system.

Class two punctuation is the most interesting. All sentence-enders are translated into either period or question mark, depending on whether the general intonation pattern desired is (M)HL, for declaratives, or (M)LH, for questions[4]. For quotes and apostrophes[5], there are two possibilities. Either the quotes enclose a single word or they enclose a phrase. If the quotes enclose a single word, then the word needs to be treated syntactically as a noun, regardless of its normal part of speech. A quoted phrase is also treatable as a noun since it can be the subject or object of a verb. There are additional complexities relating to the possibility of adjectival modification and definitization (or indefinitization) of such quoted phrases and words. These complexities should, most properly, be reviewed in a discussion of literary style rather than in a discussion of prosodic generation, and we will avoid further discussion.

---

[4] We have not extensively studied questions, but our current system uses (M)LH as the tone group for that intonation contour.

[5] Apostrophes are used like quotes except when indicating possessive and contraction.

Comma usage is even more varied and less indicative of syntactic structure than is the case with quotes. We have not yet had a chance to construct even a rudimentary algorithm for using commas. Parentheses, as well as double dashes in common orthography, set off phrasal units of one sort or another. Thus they need to be parsed as separate entities and then integrated into the sentential structure. These phrases are essentially adverbials which can occur almost freely in a sentence and are usually adjoined at the highest possible level for the constituent in question.

VII.3 Specifying the syntax

In the following figures, Figure 46 and Figure 47, we present a context free rule system for specifying the (surface) syntactic structure of sentences. Specifying this much syntactic structure enables us to generate the prosodic information necessary for

```
SENTENCE => SBAR .

        comment: The period specifies the end of the sentence.
                 This schema uses the period only to identify
                 multiply rooted trees, since the structures are
                 LISP-like.

SBAR => [WDS] (SBAR)

        comment: A bird's eye view of the bracketing
                 at this level:

                 '[ something ]      [ something else]      [ ... ].'
                 SBAR( [WDS]      SBAR( [WDS]          SBAR( ... )))

WDS => [WDS] (WDS) ! LEXEME (WDS)

        comment: The structure inside a 'something' is:

        '[ a-b]     c-e    [f-g]    [h-i]     j-k        l-m       ...
         [WDS] WDS(LEXEME [WDS]WDS([WDS]WDS(LEXEME WDS(LEXEME WDS ...

                ...  [     n-o          p-q    ] ...'
                ...  [WDS(LEXEME WDS (LEXEME))] ... ))))

                 where '-' connects the word name and part of speech.

LEXEME => (* SPECIAL) ($ STRESS) (& FREQ) (# SILENCE) (| AUDIO-COMMAND)
                          NAME PART-OF-SPEECH

AUDIO-COMMAND => terminal-string

NAME => terminal-string

PART-OF-SPEECH => N ! V ! A ! D ! C ! P ! R ! X

TERMINAL SYMBOLS : .,[,|,],*,$,&,#, word-names, audio-commands,
       X,N,V,A,D,C,P,R, numeric values for SPECIAL,STRESS,FREQ,SILENCE

        Exclusive alternatives are separated by '!'.
        Items enclosed in parentheses are optional.
```

Figure 46. CF Specifications for the prosodic control language.

PARTS OF SPEECH

N	NOUN (not a pronoun !)
V	VERB
A	ADJECTIVE OR ADVERB (including V'ed words from passives)
D	DETERMINER (NP SPECIFIER, e.g. the,a)
C	COPULA (AUXILIARY VERB when not a main verb)
P	PREPOSITIONS
R	REDUCED FORMS AND PRONOUNS (e.g. him, you're, it's, etc.)
X	CONJUNCTIONS, RELATIVIZERS, COMPLEMENTIZERS

OTHER TERMINAL SYMBOLS

.	END OF THE SENTENCE.
[	BEGINNING OF A PHRASE.
]	END OF A PHRASE.
*	indicates some special (experimental) action to be taken using the number following if desired.
$	the number following is the stress level to be assigned to the word following.
&	the number following is the peak pitch to be assigned to the word following.
#	the number following is the length of the silence to be generated at the end of the word following.

Figure 47. Explanation of terminal symbols.

reasonable sounding declarative sentences. Note that while the part of speech is explicitly shown in this schema, it could be retrieved as part of the lexical entry of the word named.

The phrase structure rules provide a means of specifying the constituent bracketing relationships within a sentence. The first rule defines the sentence to end with a period. The next level of structure for a sentence has a bracketed string, SBAR = [WDS], optionally followed by another sentence-like string. In some sense this represents an artificial notation for the concept: 'syntactic phrase with greatest scope. We actually make use of this notation when we assign our prosodic values to the syntactic structure.

The bracketed string, SBAR, is composed of a unit, WDS, which can be any major constituent at that level of the tree. Optionally, this unit can be followed by other units. Each unit is composed of a sequence of bracketed units and LEXEMES (individual lexical items). Each LEXEME can be specified in advance of the prosodic generation by using the special punctuation. This assigns a stress, frequency, or silence level which will be used in place of the one normally generated for that word[6]. The NAME is some word-name used to retrieve the lexical item from the dictionary and the PART-OF-SPEECH is retrieved as part of the entry for that name and is used to determine the neutral word stress (duration), and, along with the bracketing information, the pitch contour for the overall sentence. The SPECIAL is an option for experimental purposes. It is currently unassigned, but can have an arbitrary assignment associated with it. For example, the special might be used to cause side effects that distort the effective shape of the tree. The AUDIO-COMMAND has no effect on the prosodic manipulations performed. When present it is passed intact to the speech synthesizer, which uses the command to synchronize the synthesis with real time activities such as terminal displays.

Examples of bracketed sentences are given below (Figure 48). The first three sentences represent possible (legal) phrase structures

---

[6] Stress is more complicated than a simple replacement of preselected value for normal value. It causes side effects which distort the effective shape of the constituent structure.

1. [the north wind][and[the sun]].
2. [[the north wind][and [the sun]]].
3. [[the north wind] and [the sun]].
*4. [the north wind] and [the sun].

Figure 48. Sample bracketed strings.

for the same phrase. The fourth example is an illegal bracketing of the phrase. In this example, the word 'and' is not enclosed by bracketing at any level, therefore it is not part of any phrase and the structure is illegal. The first example shows two phrases, neither of which is embedded in the syntactic tree. In that example the word 'and' is part of the second phrase. In the second example there are two constituents embedded one level deep in the structure, and, in the third example there are three constituents embedded one level deep in the structure. Examples one and two differ in that there is no single root to the tree in example one, while example two has a root. We will see that there is a difference in prosodic assignments for these two constituent analyses.

While the particular format for specifying constituents is not important[7], it should be noted that the bracketing used for prosodic generation is not a labeled bracketing. Our theoretical tone-group elaborations made no use of information about the syntactic class of a constituent other than the information directly implied by the syntactic classes of its lowest level elements, and the actual synthesis system incorporates that idea. We are predicting that there is no prosodic difference, for example, between noun phrases and verb phrases other than those that result from the difference in internal structure and the lowest level elements in each[8] [9]. This is not to say that there is no information other than surface syntactic structure which plays a role in prosody. Our hypothesis in using this phrase structure system is twofold: a) We were principally interested in neutrally intoned sentences; b) Any perturbations to the neutral assignments can marked into the phrase structure in a relatively clean way. For example, semantic interactions which result in stress can be marked simply in terms of the '$ STRESS' parameter we have allowed. An emendation of this system will be required to specify phrase level parameterizations which may be necessary for incorporating text-view information into the prosodic assignment.

---

[7] It was designed to be general enough to be compatible with most immediate constituent analysis methods for determining surface structure. In particular, the near surface output of a syntactic generator is able to supply all the bracketing information necessary for prosodic generation.

It is conceivable that if a relative, lexically based stress (duration) assignment system were employed (as we discussed above in

---

[8] Noun-noun compounds are a possible counterexample to this claim. To maintain the original claim, all that is required is a context-sensitive assignment of prosodic features with a special clause for 'noun in the context: a following noun in the same phrase.' Then we would not say that the 'noun-noun compound phrase' needs to be labeled as such. Certainly, there is no point in simply using the standard syntactic label for the phrase, since that by itself doesn't give us enough information about what the contour assignment will be.

[9] Another possible counterexample comes from D. Klatt (personal communication), to the effect that subjects, clauses and sentences show phrase final lengthening while other phrases do not. Duration assignment may well be somewhat different from pitch assignments.

Section IV.5.1), the only part-of-speech information we would need would be the distinction between stress and unstress assigned to a word in the lexicon, rather than relying on the part of speech for stress information.

VII.4 Prosodic specification generation

After the parser is finished, a bracketed string as described above is sent to the parameter generator. The table below outlines the rules for intonation assignment. There is one intonation assignment rule for each of the phrase structure rules governing the constituent structures. While the syntax of the language specification is context free, the intonation assignment is context sensitive in order to properly handle the phrase initial upstep, phrase final falls and other phenomena mentioned above.

The assignments given below (Figure 49) refers only to neutral stress cases. Where a special stress marker is used, it distorts the tree structures so that the effect of these rules, while still largely pertinent, is obscured. Roughly, stress is generated by changing the contour from the stressed word on, increasing the duration of the stressed word, increasing the loudness of the stressed word, and placing silences around the stressed word. Stress is discussed in more detail in the theoretical part of this work.

The prosodic generator traverses the bracketed string from left to right, starting at the bottom and assigning intonation parameters to the individual words of the phrase as a function of the

```
1. <freq,step,sleng> ==> <freq,step,sleng> .

       comment: SENTENCE => SBAR .

2. <freq,step,sleng> ==> [<freq,step,sleng>] (<freq,step,sleng>)

       comment: SBAR => [WDS] (SBAR)

3. <freq,step,sleng> ==>
                <freq,step/2,sleng/sfact> (<freq-step,step,sleng>)
               ! <freq-step> <freq,step,sleng> / upstep&INITIAL-ON
               ! <freq> (<freq-step,step,sleng>) / no upstep

       comment: WDS =>  [WDS] (WDS)
                       ! LEXEME (WDS) / upstep&INITIAL-ON
                       ! LEXEME (WDS) / no upstep
```

Figure 49. Prosodic assignments corresponding to the phrase structure rules.

bracketing. Each of the branching non-terminal nodes, SENTENCE, SBAR and WDS, has a triple, <frequency,step-size,silence-length>, associated with its node. The other non-terminal, LEXEME, has only <frequency>, associated with it. Frequency here is the value, in herz, of the tone level for that unit of analysis. We have explained above how the contour assignment is defined recursively at each level of analysis. The step-size is the effective measure of depth of constituent structure. Upon entering an embedded constituent the step size is halved[10]. The <silence-length> gives the amount of silence associated with an end of constituent. The current algorithm allows us to decrease the silence length by some constant factor, labeled sfact, in the same way that we decrease the step size. Most of our experiments have avoided using the silence length at all, in order to control better the other assignments' effects on perceived intonation. Avoiding silences in neutral contours has still resulted in natural sounding sentences.

---

[10] When the step size reaches zero, it is set to a minimum of one.

The first two rules serve to propagate the initial frequency and step size through all of the topmost constituents. This propagation makes the tree structure much flatter than the phrase structure rules alone indicate. Only when going down one branch of the tree, does the program modify the current-frequency and current-step. The first expansion of rule three gives the downstepping (or upstepping if the step is negative) between sub-constituents of a non-sentential level. The second and third expansions assign actual frequencies to the lexical items (word names). As each left bracket is encountered, an INITIAL-ON flag is set to true. When a word name is encountered, its lexical (hash or quick lookup table) entry, specifically its part of speech, is retrieved. The part of speech is used to help determine the direction (up or down) of the step and the durational assignment. If the word is a minor part of speech and there is more lexical material in the same phrase and the INITIAL-ON flag is set (which means that the current word is the first one in the current phrase), then an upstep, which we represented phonologically as a 'mid to high', is called for[11]. In the case of an upstep the second constituent (which may itself be complex) will receive the high tone for the phrase. This is represented by the second expansion of rule three. We can note that in this second constituent there may also be an initial upstep so that the second word is not the recipient of the high tone. We can see this in the following example, where f is the frequency of the high tone for the phrase and s is the initial step size for that phrase. Note that the embedded constituent uses a smaller step size (s/2) than the parent constituent, due to rule two.

---

[11] Some additional complications might be needed to account for 'weight' of the phrase as determining, in part the possibility of upstep.

```
[in    [an    effort] ]
 f-s           f
 f-s   f-s/2   f
```

If the current word does not receive an upstep then the following constituent will be downstepped (by step) following the third expansion of rule three. Other parameters stored in the lexical entry for the word, such as maximum pitch and initial pitch, will be used later, for the actual imposition of the prosodic curve.

At each right bracket, the program decides on the appropriate fall for the final constituent of that bracketing. The size of the fall is proportional to the step size within that constituent[12]. When the procedure reaches the end of the current bracketed string (the current SBAR unit), and before starting on the next, it "looks back" and assigns the principal stress (full duration) to the last major part of speech: noun, verb or adjective/adverb. When the procedure reaches the end of the sentence (signaled by the "."), sentence final lengthening is generated.

---

[12] Actually, the part of the fall that is proportional to step size is added to the "normal" fall, that is assigned to every word that is not smoothed to the beginning of the following word, that is, a fall to a frequency ten per cent below the prescribed maximum.

```
1.              [the north wind] [and [the sun]].

                     SENTENCE <f,s>
                             |
                        SBAR<f,s>
            |----------------|---------------|
        <f,s>[WDS]                      SBAR<f,s>
         /    \                              |
<f-s>LEXEME WDS<f,s>                    [WDS]<f,s>
 the-D       /  \                        /     \
           /    \               <f-s>LEXEME   WDS<f,s>
     <f>LEXEME  WDS<f-s,s>       and-P            |
      north-A       |                             |
                LEXEME<f-s>                   SBAR<f,s>
                wind-N                             |
                                              [WDS]<f,s/2>
                                                 /     \
                                     <f-s/2>LEXEME  WDS<f,s/2>
                                        the-D           |
                                                    <f>LEXEME
                                                     sun-N

2.         [[the north wind] [and [the sun]]].

                       SENTENCE <f,s>
                           |
                        SBAR<f,s>
                           |
                        [WDS]<f,s>
               |-----------|-------------------|
        <f,s/2>[WDS]                       WDS<f-s,s>
            /    \                              |
<f-s/2>LEXEME   WDS<f,s/2>                 [WDS]<f-s,s/2>
 the-D             /   \                     /        \
         <f>LEXEME  WDS<f-s/2,s/2>  <f-s-s/2>LEXEME  WDS<f-s,s/2>
        north-A        |            and-P                |
                    lex<f-s/2>                       [WDS]<f-s,s/4>
                     wind-N                            /    \
                                         <f-s-s/4>LEXEME  WDS<f-s,s/4>
                                             the-D              |
                                                            <f-s>LEXEME
                                                              sun-N
```

Figure 50. Tree structures showing the assignments of prosodic parameters to bracketed strings.

The following examples (Figure 50) follow the frequency (f) and step size (s) variables through two possible assignment trees. We are showing the assignment of only these variable since silence lengths are handled similarly, falls are related to the frequency and step assignments, and duration assignment is not a function of the tree structure.

Assume that each sentence entered with <f,s>. Then we can summarize the assignments to the individual words of the sentences in the table below (Figure 51). The first noticeable difference between the two assignments is that a smaller effective step has been employed in the second example. The step is smaller because the constituents are all embedded one level deeper. The second difference between the two is that there is no downstep between the most major sub-phrases of the first sentence while there is a downstep between the most major sub-phrases of the second. Notice that we have, as a by-product of the assignment procedure induced a downstep between 'wind' and 'and' in example two, while in example one both are on the same pitch level.

VII.5 The command translator

1.	[the	north	wind]	[and	[the	sun]].
	f-s	f	f-s	f-s	f-s/2	f
2.	[[the	north	wind]	[and	[the	sun]]].
	f-s/2	f	f-s/2	f-3s/4	f-5s/4	f-s

Figure 51. Summary of prosodic assignments in sample cases.

After generating all of the parameters for frequency level, silence length (when used), fall extent and duration fraction, the information is either sent to the command translator, or is made available for experimental manipulation prior to command translation. There are two different synthesis procedures which use the translated commands to generate listenable speech. Regardless of the actual synthesis procedure used, we can regard the synthesis as being done locally, i.e. only a small part of the phonetic specification of the word being worked on and the dictionary entry for the current word is available to the procedure. The command translator calculates the duration shrinkage-factor and the frequency-multiplier. The shrinkage-factor is simply a fraction deduced from its duration-level assigned above[13]. The frequency-multiplier is calculated as the quotient of the maximum frequency of the lexical word and the pitch assigned to that word as part of the contour. The target pitch for the fall at the end of the word is calculated either: a) directly from the specification, in the case of an unusual fall (e.g. phrase final), or, b) from the beginning of the following word, if that word begins with a voiced segment, or, c) from the calculation of a normal fall to a level ten per cent below the assigned pitch level.

---

[13] If we were using an absolute, relative duration assignment, the shrinkage-factor would be multiplied by the quotient of the absolute maximum time and the duration of the lexical word.

VII.6 Prosodic contour imposition

Sequential frames of predictor coefficients are processed using the command string. Each frame has its duration shortened by the duration factor, its frequency multiplied by the appropriate fraction needed to reach the maximum pitch, and its volume multiplied to the assigned level between zero and one. When the frame of maximum pitch for the word is encountered, the incline-factor is used to impose a linear incline (fall or rise) on the rest of the current word (except that unvoiced segments are not assigned tone) so that the final pitch corresponds to the desired fall level. All words except the last are processed this way. The last word is processed similarly except that instead of having a linear fall imposed after its maximum is reached, its lexical contour is simply normalized[14] by using the frequency-multiplier for that word.

If a number of sentences are to be combined, an appropriate silence is added at the end of the last word of each sentence. A description of the actual processing as performed by the MISS machine can be found in Sanders and Benbassat (forthcoming).

---

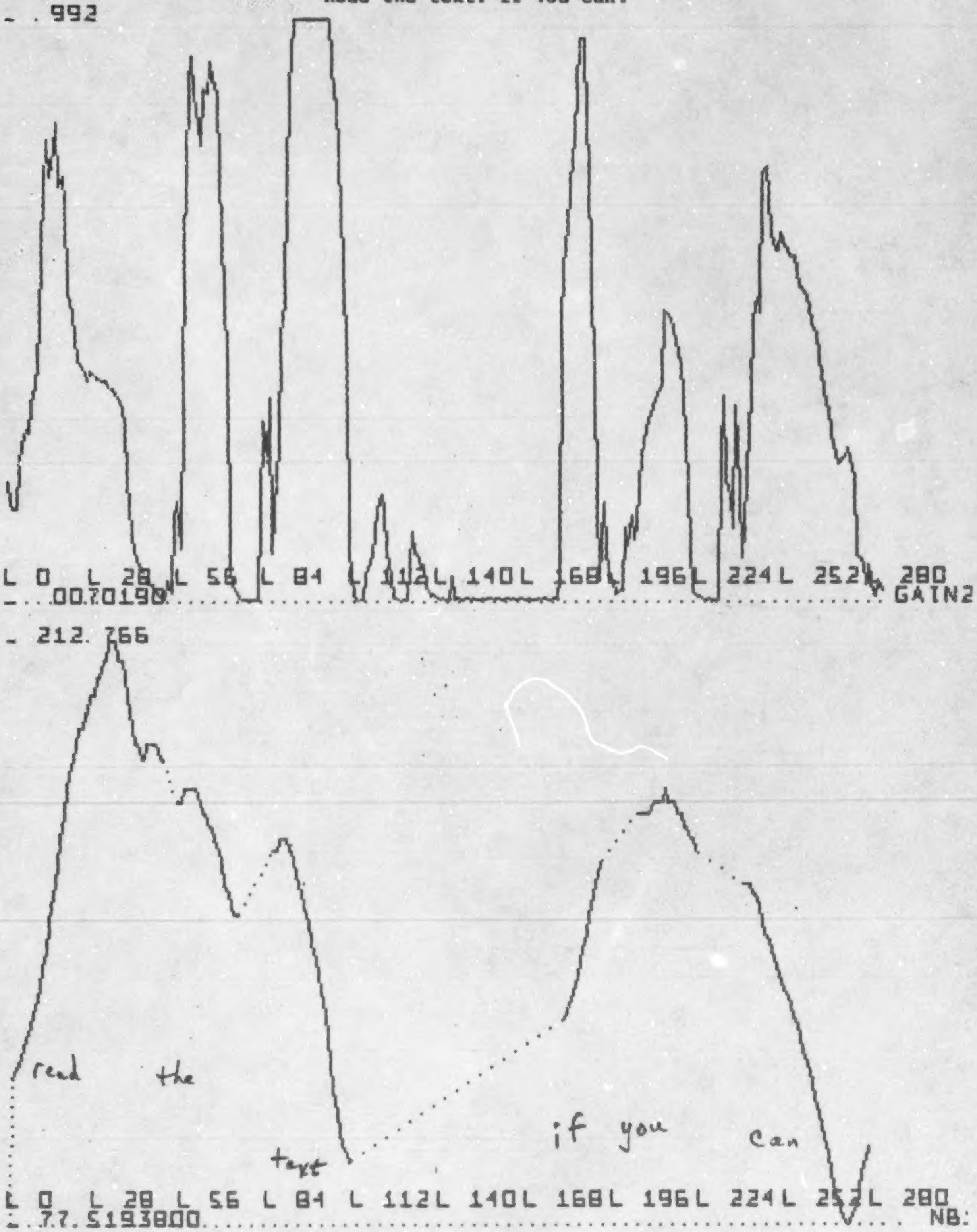
[14] We have been re-considering the effect of leaving the last word in its lexical form and now treat all the words similarly.

Appendix I

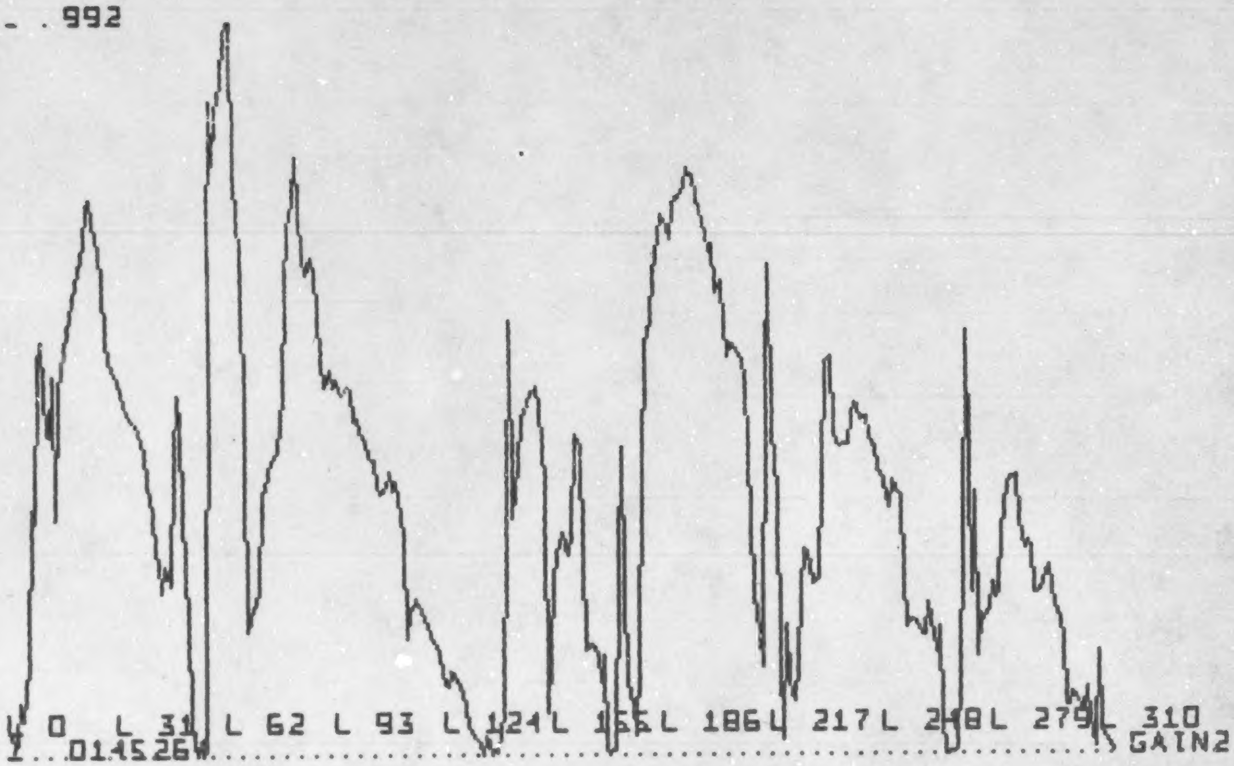
Fundamental frequency and loudness graphs

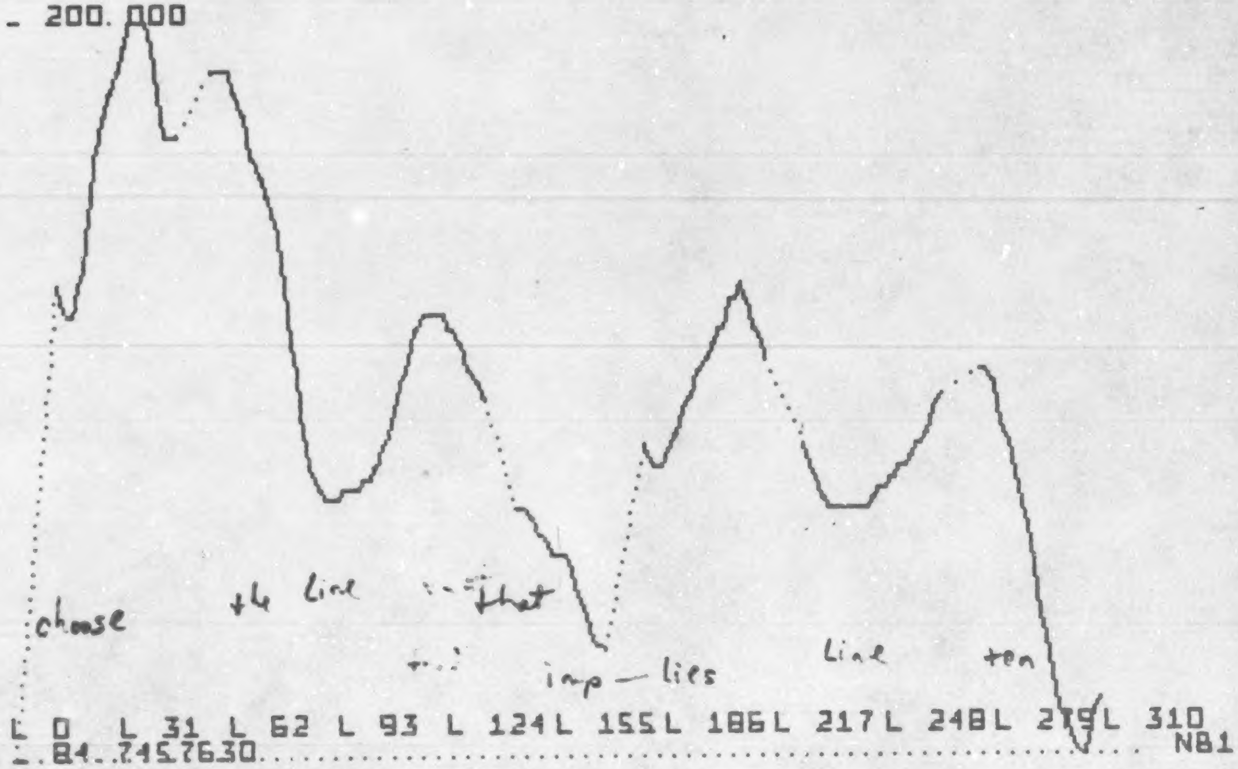
The following pages give the loudness and fundamental frequency (F_0) graphs for selected sentences. In each case, the upper graph is loudness, normalized to range from 0 to 1. The lower graph shows F_0 in hertz. The solid lines are used for voiced segments and dotted lines are interpolated between them for voiceless segments. Each F_0 is scaled from lowest pitch to highest (in that sentence). The horizontal axis gives time in hundreths of a second.

Read the text: if you can.

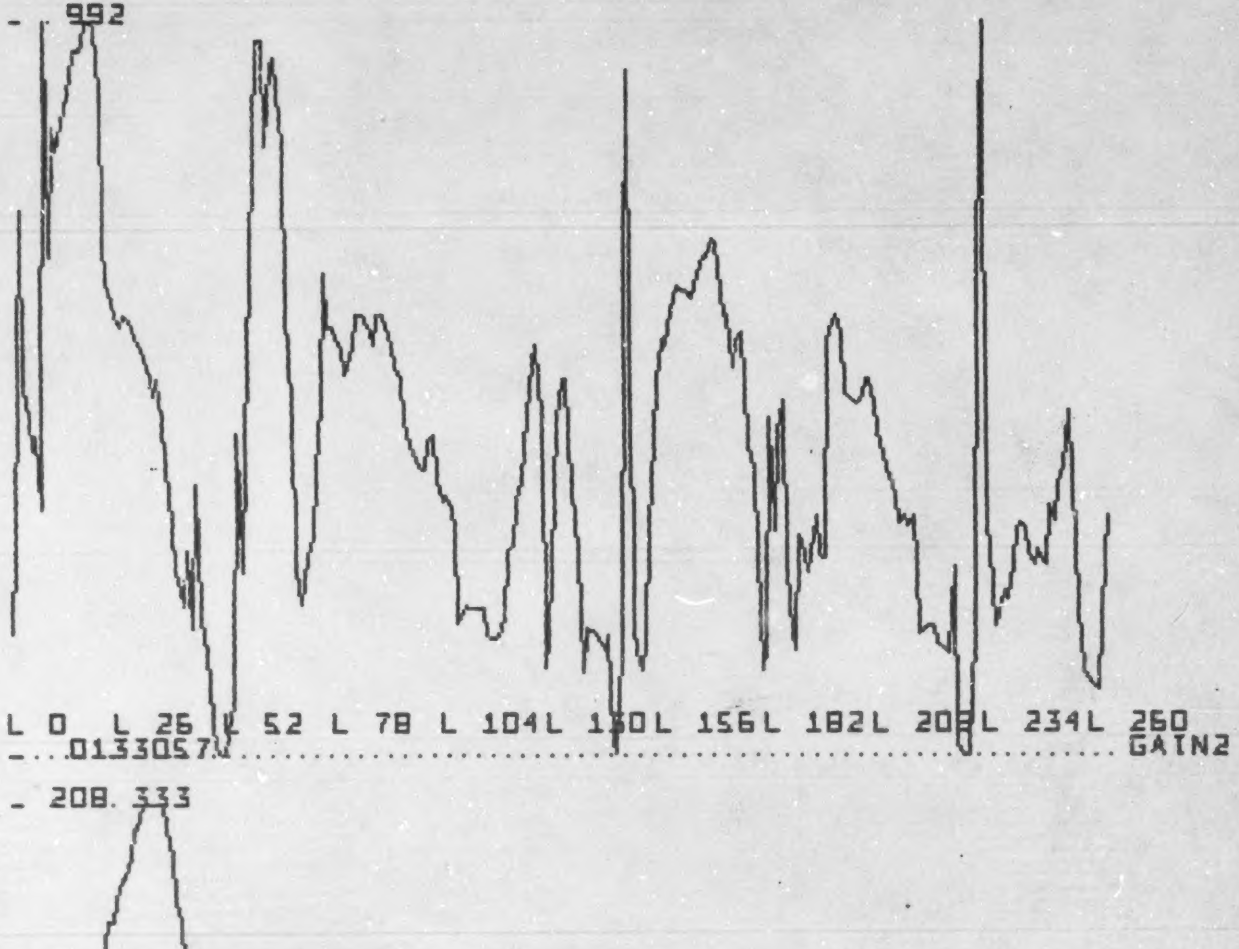


Choose the line that implies line ten.



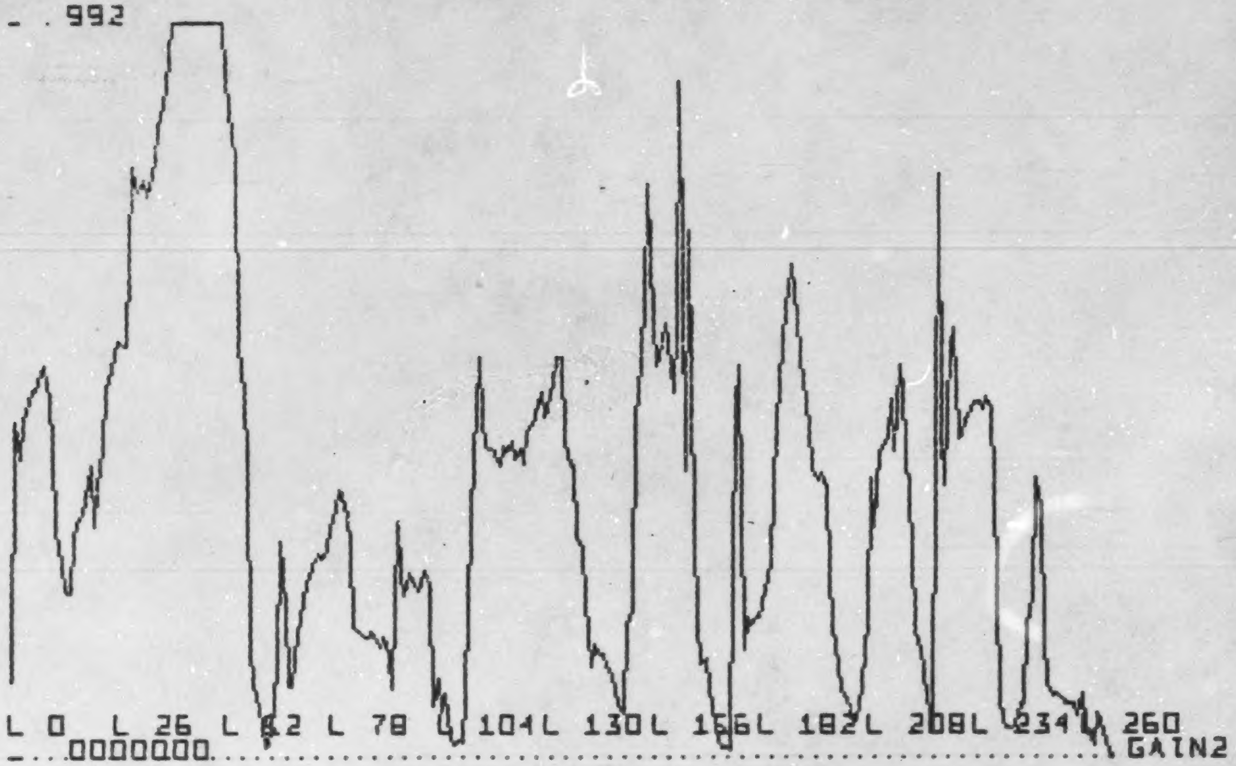


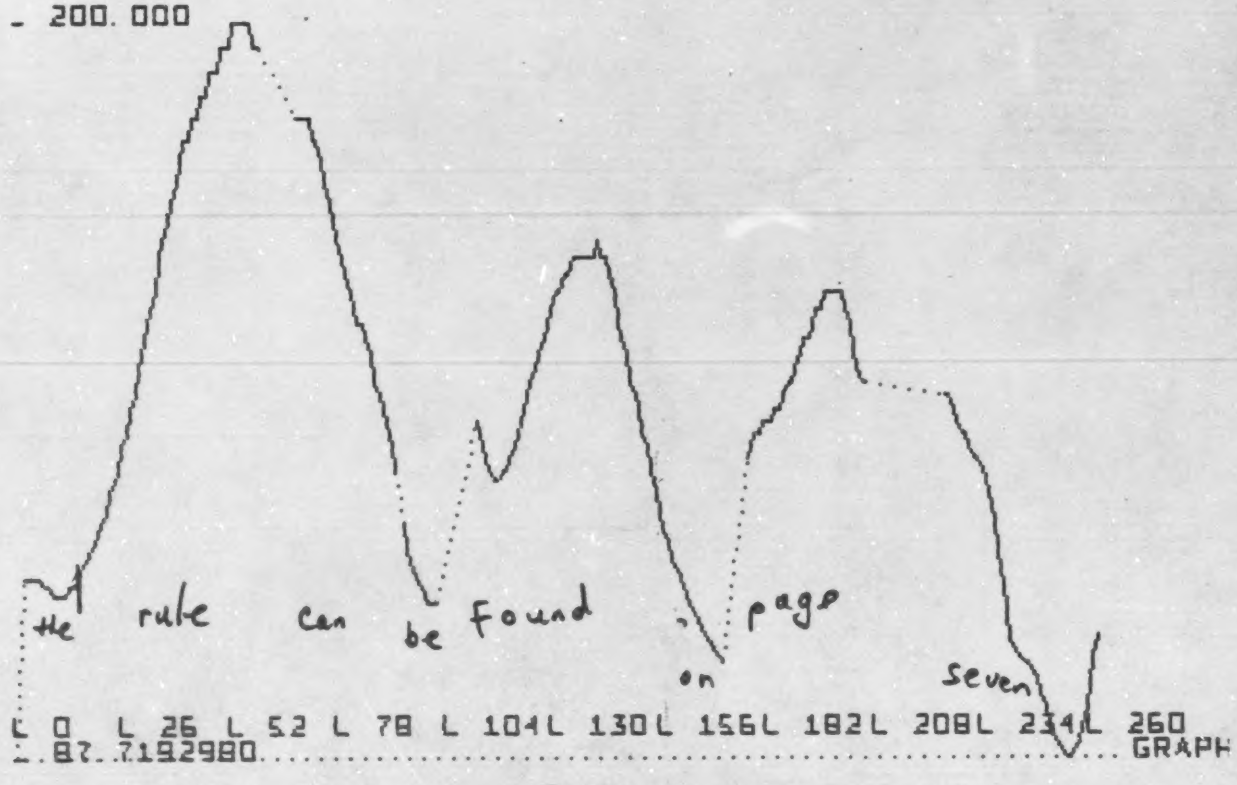
Choose the line that implies line ten.



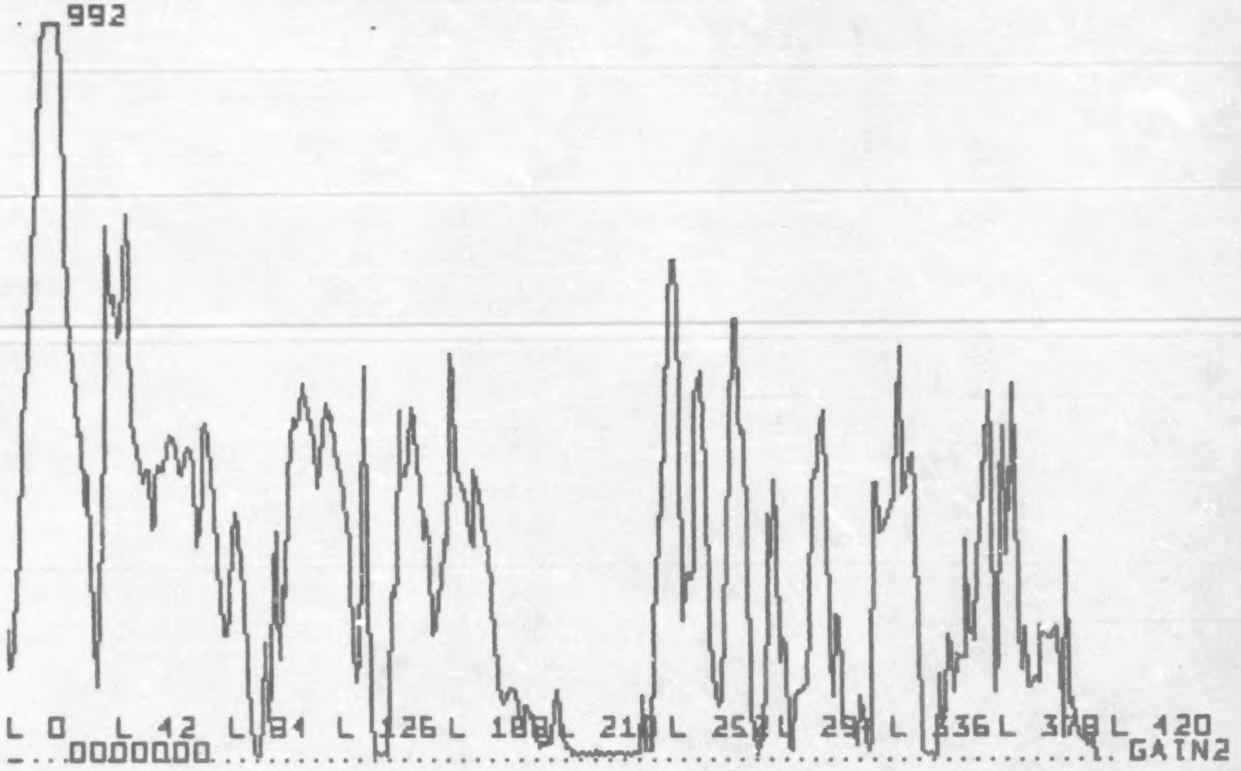
262

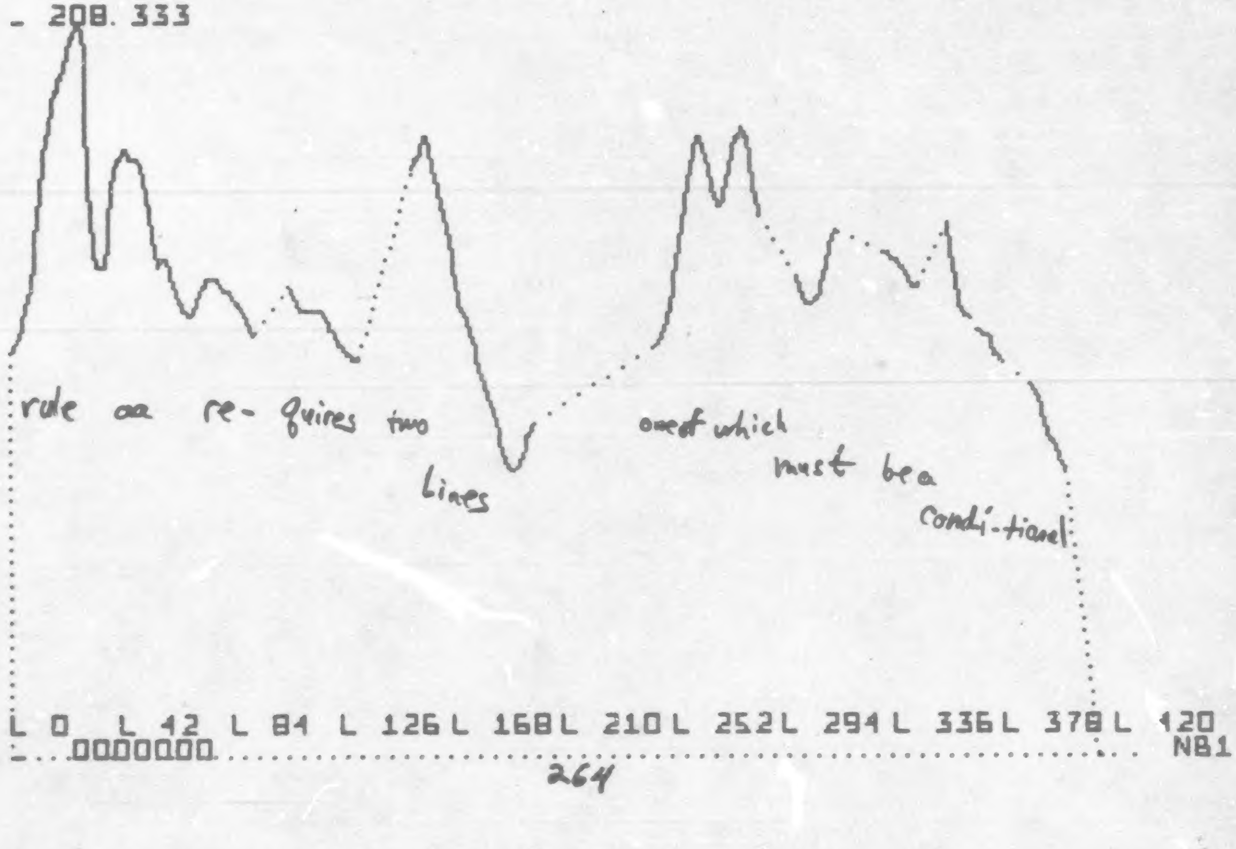
The rule can be found on page seven.



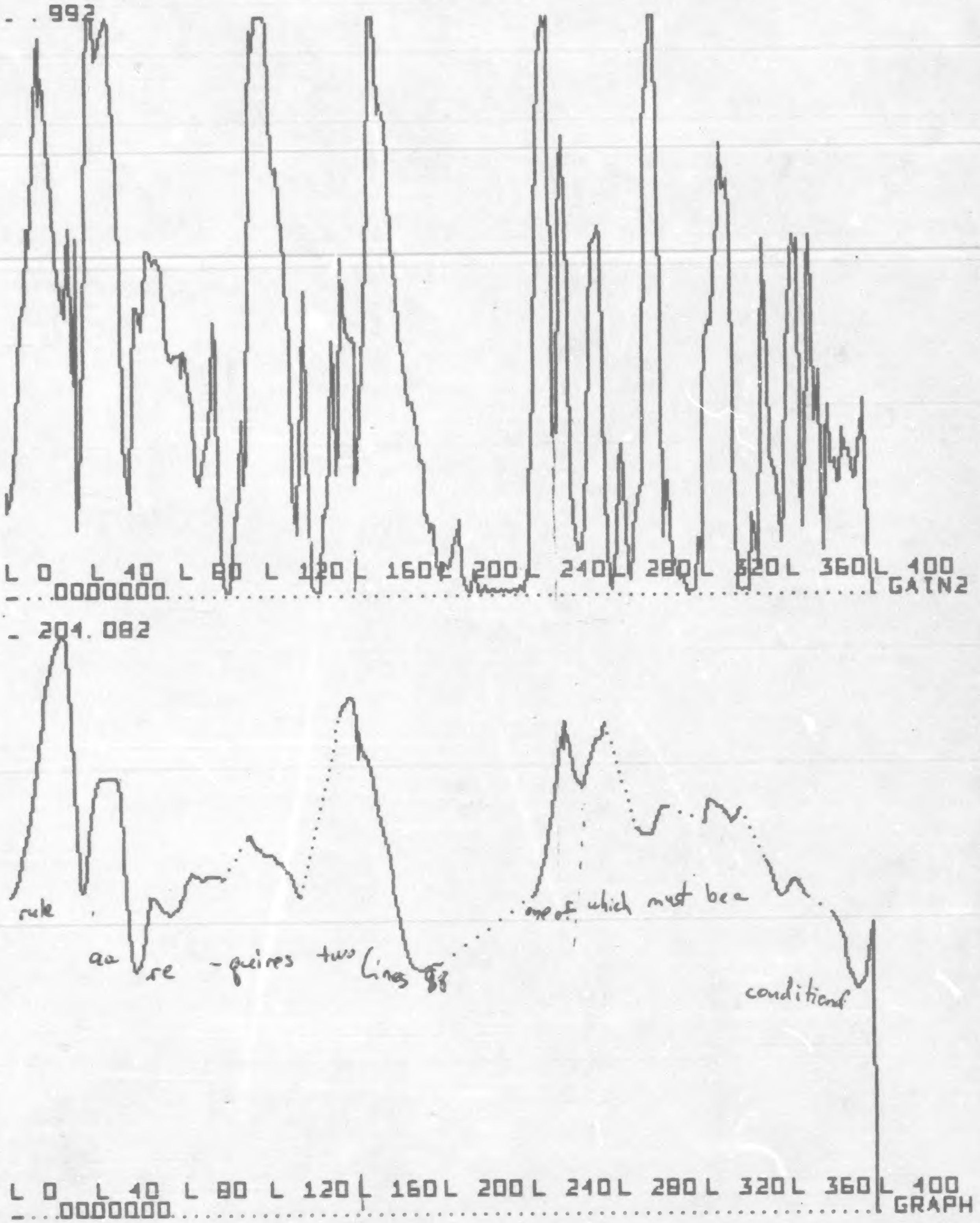


Rule aa requires two lines; one which must be a conditional.



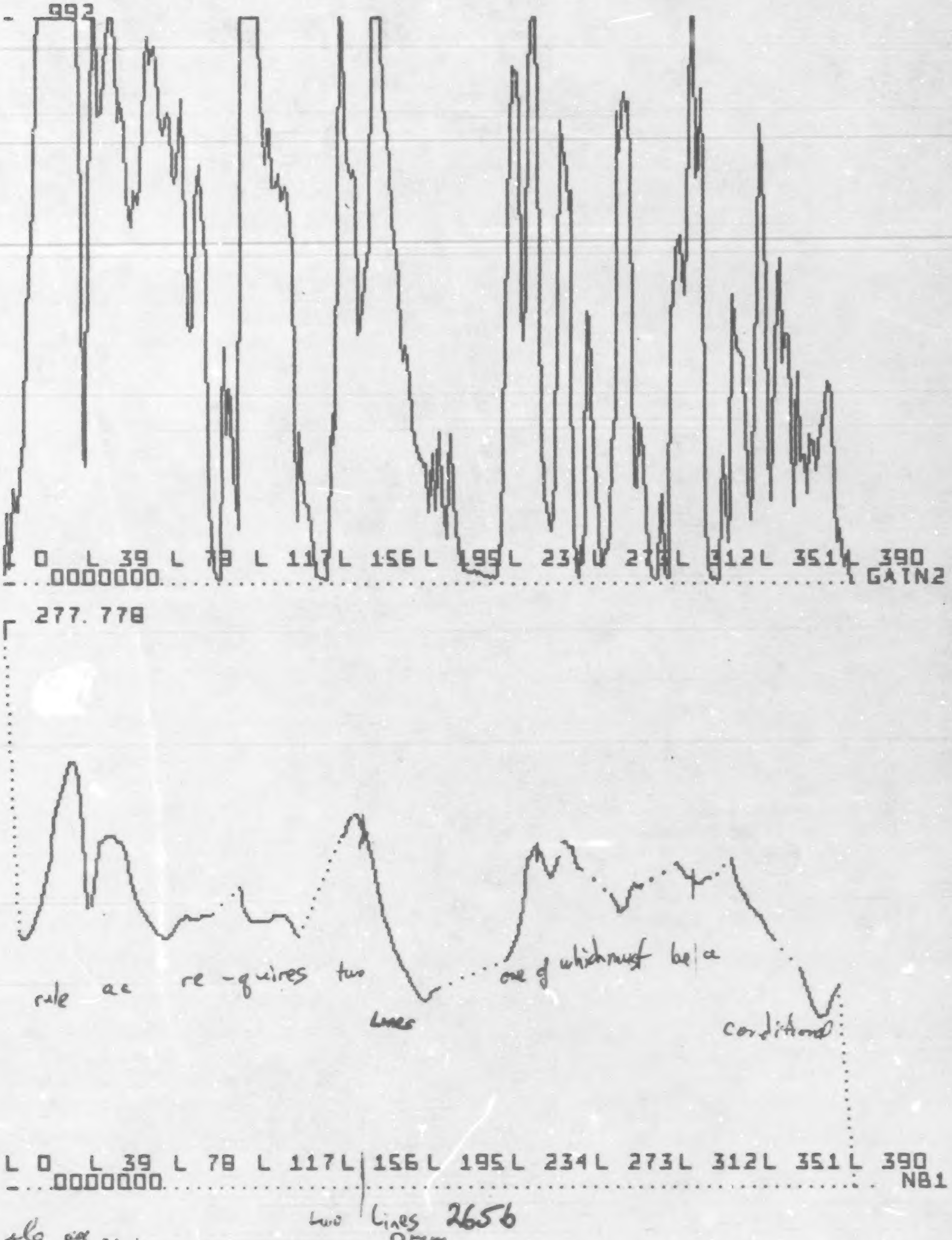


Rule aa requires two lines; one which must be a conditional.



ule aa requires two (lines 265a one of which must be a conditional

Rule aa requires two lines; one which must be a conditional.



Appendix II

The North wind and the sun

Introduction and explanation

We give in this appendix peak pitches from four separate readers and from our prosodic algorithm for the paragraph "the north wind and the sun." Several notes are required.

1) The data for this paragraph is reduced from recordings made for an introductory phonetics class by students in the class. The recordings were made in different years, hence slight variations in the texts occur. The recordings were originally made for phonetic transcription practice. All of the texts were read from a typed script and recorded on audio tape as a continuous paragraph.

2) The tapes of the recordings were all analyzed using the LPC analysis program at the Institute[15]. Timing information for the individual words in the recordings were obtained by a combination of segmenting by listening, and, pitch and loudness information from the analysis. The timing information was used, in turn, to identify the peak pitch of each word.

---

[15] Cf. Sanders and Benbassat (in prep).

3) The recordings labeled ea and sp were recorded for the same year of the course and they are identical texts, except that sp recorded the title but not her name whereas ea did the opposite, recording her name but no title.

4) The recordings labeled al and sil were done by male speakers in their mid-twenties, those labeled ea and sp by females of roughly the same age. sp is not a native speaker of English (her language is Spanish), all of the other speakers are originally from the northeastern United States, having spent from one month to several years in California. sp's Spanish accent is clearly noticeable on the recording.

5) Line numbers have been added for convenience and have no importance. Each data line is prefaced by initials indicating the speaker, as described above. m is the result of the machine processing.

6) One version of the text was arbitrarily selected as 'basic'. We have tried to organize the recordings so that all versions of the same part of the text are given under the same columns. When the differences are very great a separate line of text is included above the recordings it refers to. In line 7, the al and sil recordings use the first line of text and the ea and sp use the second (shorter) line.

Minor differences from the basic text are enclosed in angled brackets ('<' and '>'). In line 8 the word 'off' occurs in two positions in the different readings. For sil the phrase reads "take off his cloak"; for all other readers it reads "take his cloak off." The period at the end of line 8 indicates that that was the end of the sentence for ea and sp but not for al and sil.

7) All numbers refer to peak fundamental frequencies in hertz. Where no voiced segment was associated with a word a dash ('-') replaces the fundamental frequency. When, due to text differences, a word does not appear in the recording, no mark is put in that column. Parenthesized numbers are approximations. In line 3 the processing number is the peak pitch if the terminal rise due to the speaker's chuckle is eliminated. The parenthesized numbers in lines 12 and 16 indicate uncertainty as to the correctness of the pitch value recorded due to errors in the pitch detection and analysis programs. In line 17, the parenthesized value at the end of the line is an alternative for the word 'two' adjusting for some possible pitch error in the analysis.

The text

Linguistics two-fifteen[16]

Foundations of phonetic and phonological analysis[17]

f-name l-name[18]

The north wind and the sun[19]

The north wind and the sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler[20] take his cloak off[21] should be considered stronger than the other[22]. Then the north wind[23] blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the north wind gave up the attempt[24]. Then the sun shone out warmly, and immediately the traveler took off his cloak. And so the[25] north wind was obliged to confess that the sun was the stronger of the two.

---

[16] Only read by al.

[17] Only read by al.

[18] Three readers read their own names.

[19] Only read by al and sp.

[20] Alternate version: "They decided that whoever could make him"

[21] Alternate version: "take off his cloak"

[22] Alternate version: "considered the stronger."

[23] Alternate version: "The north wind then"

[24] Alternate version: "could, but the traveler only folded his cloak more tightly about him."

[25] Alternate version: "The"

The data

1. Linguistics two-fifteen.

al 159 133 122

m 155 135 115

2. Foundations of phonetic and phonological analysis.

al 152 100 149 94 133 119

m 155 142 145 130 135 130

3. Arvin Levine.

al 123 116(103)

Bill Sanders

sil 172 147

Elaine Andersen /chuckle/

ea 192 (175)233

m 155 135

4.	The	north	wind	and	the	sun.
al	104	161	143	130	115	123
sp	185	263	270	185	169	169
m	150	155	145	130	132	135
m	145	155	135	145	150	155

5.	The	north	wind	and	the	sun	were	disputing	which	was	the	stronger,
al	114	196	152	135	114	109	103	156	122	130	132	108
sil	153	256	166	149	147	166	136	196	163	172	163	169
ea	227	303	263	222	-	222	193	227	196	189	182	217
sp	200	294	250	169	159	196	238	213	263	200	185	175
m	152	155	150	142	144	145	130	135	145	150	154	155
m	152	155	150	142	144	145	132	135	122	124	124	125

6.	when	a	traveler	came	along	wrapped	in	a	warm	cloak.
al	122	95	145	143	125	116	101	96	130	122
sil	142	135	208	144	133	151	147	131	136	158
ea	196	185	250	217	200	208	189	182	185	182
sp	189	238	250	200	182	278	179	217	189	164
m	152	153	153	150	149	148	147	147	147	147
m	155	150	155	135	125	115	102	104	105	103

7.	They	agreed	that	the	one	who	first	succeeded	in	making	the	traveler
al	121	145	156	123	122	112	175	143	149	156	141	122
sil	200	238	129	138	163	138	212	200	136	156	125	131
m	145	155	145	150	155	145	155	135	133	130	112	115
m	145	155	155	130	135	124	125	124	119	120	118	118

	They	decided	that	whoever	could	make	him
ea	217	256	238	303	278	250	227
sp	294	294	167	263	208	182	169
m	145	155	144	145	139	140	137

8.	take	<off>	his	cloak	off	should	be	considered	<the>	stronger	<.>
al	128		135	123	104	109	103	119		125	
sil	133	153	126	117		156	149	164		175	
ea	217		110	213	192	238	233	208	156	179	.
sp	250		189	182	156	238	233	208	156	179	.
m	105		134	135	133	129	129	130		130	
m	117		117	117	117	145	150	155		135	

9.	than	the	other.
al	115	110	125
sil	135	151	158
m	130	132	135
m	122	124	125

10.	Then	the	north	wind	<then>	blew	as	hard	as	he	could,
al	175	143	152	116		130	99	135	123	122	91
sil	217	196	163	135		158	140	175	172	178	158
ea		196	244	227	222	204	208	222	208	238	217
sp		169	238	278	238	256	227	208	169	179	147
m	155	150	155	145		135	145	155	130	132	135
m	155	150	155	145		135	122	125	120	114	115
m		145	155	135	145	155	145	155	130	132	134

11.	but	the	more	he	blew	the	more	closely	did	the	traveler
al	122	101	154	152	128	127	125	147	139	101	132
sil	131	123	172	181	161	121	142	196	161	149	158
m	145	150	155	145	155	144	145	143	139	140	140
m	155	150	155	130	135	150	155	145	130	134	135

	but	the	traveler	only	drew	his	cloak	more	tightly	about	him.
ea	217	213	250	213	208	222	217	208	222	192	172
sp	172	156	250	227	217	208	192	238	222	161	154
m	155	152	155	145	135	124	125	123	121	120	155

12.	fold	his	cloak	around	him;	and	at	last	the	north	wind
al	141	(159)	132	125	118	102	100	127	97	139	122
sil	161	149	158	126	113	136	133	151	129	163	149
m	139	139	139	139	155	155	145	155	150	155	145
m	130	128	127	125	123	155	130	135	150	155	145

13.	gave	up	the	attempt.
al	122	114	102	123

sil	142	129	138	175
m	135	125	145	155
m	135	125	112	115

14.	Then	the	sun	shone	out	warmly,	and	immediately	the	traveller	took
al	175	119	145	130	130	125	103	147	-	122	112
sil	227	212	185	166	156	158	133	196	126	144	135
ea	270	278	250	217	182	189	208	213	118	200	222
sp	250	167	189	256	233	196	175	286	159	222	238
m	155	150	155	135	122	125	145	155	132	135	125
m	155	150	155	135	130	125	145	155	150	155	135

15.	off	his	cloak.
al	114	118	105
sil	144	136	149
ea	182	192	196
sp	192	172	161
m	119	119	120
m	130	122	125

16.	And	so	the	north	wind	was	obliged	to	confess
al	(58)	137	127	137	115	-	125	127	143
sil	149	163	149	188	172	129	133	-	153
ea			204	250	115	179	204	208	200
sp			161	213	213	196	217	189	208
m	155	145	152	155	150	144	145	139	140
m	150	155	132	135	130	122	125	119	120

17.	that	the	sun	was	the	stronger	of	the	two.
al	130	99	152	111	109	125	127	125	125(92)
sil	-	151	149	138	147	142	123	-	158
ea	196	200	233	200	192	208	175	172	204
sp	278	250	217	179	263	238	189	143	132
m	145	152	155	142	144	145	142	143	143
m	115	132	135	129	130	130	129	129	129

Appendix III

Text of the lesson with comments

The lesson shown below is taken from an elementary course in logic given for academic credit at Stanford University, Autumn 1976. The course consisting of thirty lessons is completely conducted using computer terminals equipped with audio playback facilities[26]. The course is designed so that the student is presented with a discussion of the material using the audio, coordinated with displays on the terminal screen. We have reproduced below the audio text of one lesson of the course. Display commands, which coordinate the visual presentation at a computer terminal with the spoken text, have been edited out since they are not critical to this examination and their absence does not impede understanding of the text. Notice that some of the sections below are clearly oriented towards visual display of information, while others are largely verbally oriented. This difference results from the nature of the material presented, but for the most part, our analysis can be independent of such considerations.

The following conventions will be in effect for the lesson given in this appendix:

1) Each lesson is divided into "exercises" which are the largest units examined by our analysis. The

---

[26] A further discussion of this course, and the uses of audio material for computer aided instruction can be found in Sanders and Laddaga (1976).

beginning of each exercise has a heading which gives the lesson number (offset by 400) and the exercise (or paragraph) number, e.g. "lesson 402 2" is lesson 2 in the course, exercise 2. Since these exercises are taken from actual lessons in use at Stanford University, we have left them essentially unedited. Three features of the lesson texts should be noted.

1) Capitalization of entire words is freely used in the text, generally as an indication of emphasis or stress.

2) Special word stress is indicated by a dollar sign followed by a number, usually 1, e.g. "$1 hello" in the written text indicates a special heavy stress on the spoken word "hello". $3 is less stress and more isolation (See the main text Section VII.3), it is sometimes used to generate minor pauses in the spoken text.

3) Parentheses are often used instead of commas in the text. This usage developed because of the normal ambiguity of comma usage in English.

The parenthesis sets off major phrasal units.

2) The right hand column contains the analysis of the lesson. Indentation is utilized to indicate which elements are contained in which larger element. The broadest analysis sub-divisions are underlined and capitalized. Letters and numbers in parentheses are labels for intra-text reference. Footnotes are indicated by asterisks (*).

```
                      lesson 402 1
[
A
 We said in the last lesson that the variables, P, Q, R, etc, are used
        to name declarative sentences.
B
 Declarative sentences have a truth $3 value that $3 is they can be
        either TRUE or FALSE.
C
 This is $1 not true for)( $3 questions  $3 commands  $3 names  or
        $3 predicates . Those $1 aren't really true or $3 false
        But declarative sentences $1 do have a truth value.
] [ [
D
 For example, suppose we let R be the declarative sentence, (Snow is
        white).
E
 And we let Q be the declarative sentence (Grass is green).
] [
F
 Then R AND Q would be the $3 sentence) ( Snow is white $1 and grass is
        green.
G
 R OR Q would be the $3 sentence)( Snow is white $1 or grass is green.
H
 IF R THEN Q would be the $3 sentence)( If snow is white, then grass is
        green.
I
 And finally, NOT R would be the $3 sentence )( It is not the case that
        snow is white.
] ]
```

[
A
The sentential connectives are the terms of a language which we may use to join sentences together to form more complex ones.

B
As we have noted before, the $1 meaning of a connective, as with many other (if not most) words in English is usually not very clear.

C
All kinds of different meanings may be $3 given depending upon the context of the connective.

D
Many times the particular sentence will be ambiguous, or even meaningless.

E
The English $3 connective $1 and is probably the least subject to variation, but there are some cases where even $1 and is not as explicit as we want.

] [
F
Here is an example which shows how the meaning can vary with the context. The sentence $3 is)(Horace went upstairs, got into bed, and took off his skis.

[
G
You probably feel that this sentence is a bit peculiar for it would normally mean that Horace had taken off his skis $1 after getting into bed, which might not be the best way to do it.

H
That is because in this context, the connective (and) gives us two types of information.

] [
I
One thing it tells us is that the separate sentences.

[
J
$3 1 (Horace went upstairs).

K
$3 2 (Horace got into bed).

L
$3 3 (Horace took off his skis).

]
M
are all true.

]
N
The second thing which we get by having (and) connect these three sentences is that we find out something about $1 WHEN these events occur.

O
Thus we would want to say that the sentence would be false if Horace

took off his skis $1 before he got into bed, even if the parts 1, 2, and 3 alone were each true.

] [

P

Right now we are concerned with just the truth functional sense of the connectives.

Q

We will give a definition for each of the connectives which avoids the two problems of ambiguity and dependence on context.

R

These will be truth $3 functional that is, the truth value of the whole sentence will depend ONLY on the truth value of the component sentences.

]

S

For example if we know the truth value of P and of Q then we want to be able to tell $1 exactly what the truth value of any sentence using P and Q with the connectives will be.

T

Any further connotations of the connectives which are lost in this approach we can express within our system in another $3 way and the logical structure is much more clear.

lesson 402 3

[

A

The first connective we'll look at in detail is the simplest $3 one negation.

[

B

Suppose that we consider a sentence (Q), and the negation of that sentence, (NOT Q).

C

$1 Remember that here we are concerned only with the truth functional sense of our connectives.

]

D

So if we know the truth value of Q, then we can tell the truth value of NOT Q.

] [

E

Thus if Q is a true sentence, such as (Snow is white) then NOT Q is the false sentence, (It is not the case that snow is white).

F

Similarly, if Q is a false sentence, such as (Snow is black) then NOT Q will be the true sentence, (It is not the case that snow is black.)

] [

G

Some of the things we've talked $3 about are slightly different from common usage.

[

H

We translate the $3 sentences (Snow is white), and (Snow is not $3 white) into symbols $3 as $3 P and as NOT P.

I

But when we translate the symbolic $3 sentence)(NOT P back into English, we prefer to use The $3 sentence (It is not the case that snow is white).

] [

J

This is for two reasons. $1 One is that the phrase (it is not the case that) parallels the use of our $3 symbols while the English use of the term (not) $1 within a sentence, doesn't parallel it very well.

K

The other reason it is preferred is that in many contexts, it is less ambiguous than the term $1 (not).

]]

L

If you want to have more detail on this topic, or any of the similar issues brought out by the other connectives, refer to Patrick Suppes' (Introduction to Logic).

[[

A

What we have just talked about can be put into a convenient form for reference.

B

This is called a truth table, and quickly shows the explicit definition of our connective.

] [

C

On the right, the truth table is headed by a sample sentence containing the connective, and on the left, by a list of the component $3 sentences. In this case, only one $3 variable Q.

[

D

The columns under the separate variables give all the different combinations of truth values which the various variables can have.

E

The column under the whole sentence gives the truth value which the component sentence $1 must have, given the values of the variables on that row.

]]

F

In the case of negation, we only have one variable to worry about, so the table is not very complex, and rather easy to remember.

]

G

If, however, you should need to refer back to any of the truth tables that we are presenting, remember that you can use Browse Mode.

lesson 402 5

[[

A

Let's see what truth value a conjunction has in different circumstances.

B

The easiest way to do this is to give the truth table for conjunction Pay close attention to examples given below to see where the values are taken from.

] [

C

If both the conjuncts, P and Q, are each true, then the conjunction P $1 and Q is true.

D

If the first conjunct of the sentence is true and the second conjunct is false, then the conjunction is false.

E

If, on the other hand, the first conjunct of the sentence is false and the second is true, then the conjunction is also false.

F

Finally, if both conjuncts of the sentence are false then again the conjunction is false.

]]

```
                              lesson 402 6
 A
  Unfortunately, not all of the connectives are as straightforward as
         negation and conjunction.
 [
 B
  There are two reasonable ways of interpreting the English $1 OR as a
         $1 truth functional connective.
 C
  This gives us two different connectives, which are called the
         $1 inclusive or, and the $1 exclusive or.
 D
  The only difference between them is that for an exclusive $1 or,
         if both disjuncts are true then the sentence is $1 false.
         While for an inclusive $1 or if both disjuncts are true, then
         the sentence is $1 true.
 ] [
 E
  In everyday language, sometimes we use exclusive or, as $3 in
         (Marvin will either go to the movies or do his homework).
 F
  While other times we use an inclusive[ sense of $1 or, as $3 in
         (Either Bob is fatter than Mark, or he is fatter than Mike).
 G
  Most of the time, our use of $1 or is ambiguous between these two.
 ] [
 H
  In everyday language there are phrases which make it clear what we
         want to say.  One way we show that it is definitely the
         exclusive sense is by adding the phrase (but not both).
 I
  A common way of showing the inclusive sense is by using the phrase
         (and or).
 [
 J
  But in our $1 logical system we want to settle on only $1 one meaning
         for any connective.
 K
  It turns out to be simpler to use the inclusive $1 or instead of the
         $1 exclusive, so from now on, whenever we refer to
         $1 disjunction or use the symbols V and $3 OR  we will mean
         the inclusive sense.
 L
  As you'll see, we don't lose anything by this, since we can use a
         combination of connectives to express the exclusive sense.
 ] ]
```

lesson 402 7

```
[
A
 Now let's look at the truth functional definition for a disjunction.
     Remember that we'll always be referring to the $1 inclusive $1 or
B
 We do this again by giving the truth table for disjunction.
         Pay $1 close attention to examples given below to see where
         the values are taken from.
] [
C
 If both the disjuncts, P and Q, are each true, then the disjunction
         P $1 or Q is true.
D
 If the first disjunct of the sentence is true and the second disjunct
         is false, then what is the value of the disjunction?
E
 If, on the other hand, the first disjunct of the sentence is false and
         the second is true, then what is the value of the disjunction?
F
 Finally, where both disjuncts of the sentence are false, then what is
         the value of the disjunction?
]
```

[

A

We'll look now at the connective (if then).

B

Like the case of disjunction, the truth value of a conditional is not always clear in English.

] [

C

Consider the $3 sentence If Humbert loves Lolita, then Lolita loves Humbert.

D

There are four cases to consider.

[

E

When Humbert $1 does love Lolita and Lolita $1 doesn't love Humbert, then it's clear that ($3 1) is $1 false.

F

And if Humbert loves Lolita and Lolita $1 does love Humbert, then it's pretty clear that ($3 1) is $1 true.

]

G

Cases like these, where the antecedent is true, are the ones that we are usually concerned with in our everyday speech.

] [

H

We run into problems in trying to get a reasonable truth functional definition where the antecedent is $1 false.

I

For example, when Humbert $1 doesn't love Lolita, it's not at all clear from normal usage, what the truth value of ($3 1) would be.

] [

J

In mathematics and logic, a conditional is taken to be true when the antecedent is $3 false so we follow this approach.

K

In our example, if Humbert doesn't love Lolita, then we agree that ($3 1) is true, whether or not Lolita loves Humbert.

]]

L

The reasons for our following this approach really go outside what we want to cover here, but if this usage seems unreasonable to $3 you or if you are interested in this topic, you should refer to Suppes' INTRODUCTION TO LOGIC, chapter 1, section 3.

lesson 402 9

[

A

Since we went over all the values in the last exercise, we'll be fairly quick with the truth table for this connective.

B

But still pay close attention to examples given below to see where the values are taken from.

] [

C

If both the antecedent, P and the consequent Q, are each true, then what is the value of the conditional (if P then Q)?

D

If the antecedent of the sentence is true and the consequent is false then what is the value of the conditional?

E

If, on the other hand, the antecedent of the sentence is false and the consequent is true, then what is the value of the conditional?

F

Finally, if both the antecedent and the consequent of the sentence are false, then what is the value of the conditional?

]

```
                          lesson 402 10
A
 Just in case some of what we have talked about has slipped by you, we
        will give a brief review of the connectives.
[
B
 The negation of a sentence)( is $1 true when the sentence is $1 false
        $1 and  is $1 false when the sentence is $1 true.
C
 A conjunction of two sentences)( is $1 true when both sentences are
        $1 true, $3 and is $1 false when either sentence is $1 false.
D
 A disjunction of two sentences)( is $1 true when either sentence is
        $1 true, $3 and is $1 false when both sentences are $1 false.
E
 A conditional with two sentences)( is $1 true when $3 either
        (the antecedent is $1 false ) or (the consequent is $1 true)
        $3 and)( is $1 false when both
        ( the antecedent is $1 true) and (the consequent is $1 false).
]
F
 When you go on, you will get a complete list of all the connectives.
```

lesson 402 11

```
[
A
 Here is the first problem using truth tables.
B
 This problem will be pretty long, so remember that if you log out in
        the middle, you will have to start over at the beginning.
C
 We'll be trying to find out the value of the expression on the right,
        given the values of the atomic sentences on the left.
] [
D
 We'll do the first case $1 for you, and then you'll be on your own.
E
 We'll start you with table of all the possible combinations of values
        for the variables on the right.
F
 The first step for you to do is to repeat the values underneath the
        variables in the $1 sentence.
] [
G
Now we will begin to figure out what the value of the whole sentence is.
[
H
 You first look for some small part of the sentence that you can use
        your knowledge of connectives $3 on  $1 so you start from the
        inside $3 out.
I
 You look first at the innermost connective, $1 NOT.
] [
J
 Since Q is true, NOT Q must be false, so we put an F underneath the
        connective $1 NOT.
K
 Now we look for the next outermost connective, which is $1 AND.
]
L
 We've found that NOT Q is false, and we already know that P is true,
 so since a conjunction is false when either conjunct is false, then
        P AND ( NOT Q ) must be $3 false  $1 so we put an F underneath
        the connective $1 AND.
] [
M
 You will now get to fill in the rest of the table by yourself.
        Remember how we went about doing this first row, and if you
        get stuck, use Browse Mode to refer back to the truth tables
        for the connectives.
[
N
 We will speed things up a bit by referring to the columns of truth
        values by the letters ($1 a), $1 b, $1 c, and $1 d.
```

O

Where do the next entries in the truth table go? Type two letters from the letters (a,) b, c, or d.

]] [

P

We've now finished the problem, and can erase all but the value of the $3 sentence P and (NOT Q).

[

Q

You have now discovered that the $3 sentence P and (NOT Q) is only true when P is true and Q is false.

R

In every other case, the sentence is false.

]]

S

Now that you've managed to work your way through this problem the next ones should be no trouble.

T

If they do give you trouble remember that you can review the beginning of this problem.

References

1. Abelson, R. The structure of belief systems, in R. C. Schank and K. M. Colby (eds.), Computer models of thought and language, W. H. Freeman: San Francisco, pp. 287-340, 1973.

2. Akmajian, A., and T. Wasow, The constituent structure of VP and AUX and the position of the verb Be, Linguistic Analysis, 1.3, 205-246, 1975.

3. Allen, J. Speech synthesis from unrestricted text, J. L. Flanagan and L. R. Rabiner (eds.), Speech synthesis, Stroudsburg, Pa.: Dowden, Hutchinson, and Ross, 1973.

4. Barnhart, C. L. and J. Stein, The American college dictionary, Random House, New York. 1967.

5. Berman, A. and M. Szamosi, Observations on sentence stress, Language 48.2, pp. 304-325, 1972.

6. Bernstein, L. The Unanswered Question: Six Talks at Harvard (The Charles Eliot Norton Lectures, 1973), Harvard University Press, Cambridge, Mass. 1976.

7. Bolinger, D. L. Forms of English: Accent, morpheme, order. I. Abe and T. Kanekiyo (eds.), Cambridge, Mass.: Harvard University Press, 1965.

8. Bolinger, D. L. Stress is predictable (if you're a mind reader), Language 48.633-45, 1972.

9. Bresnan, J. W. Sentence stress and syntactic transformations, Language 47.2, pp. 257-281, 1971.

10. Bresnan, J. W. Stress and syntax: a reply, Language 48.2, pp. 326-342, 1972.

11. Chomsky, N. A., and M. Halle, The sound pattern of English, New York: Harper and Row, 1968.

12. Chomsky, N. A. Aspects of the Theory of Syntax, M.I.T.: Press, 1965.

13. Chomsky, N. A. Reflections on Language, New York: Pantheon Books, 1975.

14. Chomsky, N. A. On wh- movement, in P. Culicover, T. Wasow and A. Akmajian (eds.) Formal Syntax, Academic Press, 1977.

15. Crowder, R. G. and J. Morton, Pre-categorial acoustic storage (PAS), Perception and Psychophysics 5.365-373, 1969.

16. Crystal, D. Prosodic Systems and Intonation in English, Cambridge University Press, 1969.

17. Darwin, C. J. On the dynamic use of prosody in speech perception, Status Report on Speech Research, Haskins Laboratories, New York: SR-42/43 pp. 103-115, 1975.

18. Fry, D. B. Experiments in the perception of stress, Language and Speech I, pp. 126-152, 1958.

19. Gaitenby, J. H. The elastic word, Status Report on Speech Research, Haskins Laboratories, New York: SR-2, 3.1-3.12, 1965.

20. Gaitenby, J. H., G. N. Sholes, and G. M. Kuhn, Word and phrase stress by rule for a reading machine, Status Report on Speech Research, Haskins Laboratories, New York: SR-29/30, pp. 105-110, 1972.

21. Gaitenby, J. H. Stress and the elastic syllable: An acoustic method for delineating lexical stress patterns in connected speech, Status Report on Speech Research, Haskins Laboratories, New York: SR-41, pp. 137-152, 1975.

22. Goldman-Eisler, F. Pauses, clauses, sentences, Language and Speech 15.114-121, 1972.

23. Goldsmith, J. English as a tone language, unpublished M.I.T. ditto, 1974.

24. Goldsmith, J. An overview of autosegmental phonology, Linguistic Analysis, 2.1, pp.23-68, 1975.

25. Halliday, M. A. K. Explorations in the functions of language. London: Edward Arnold, 1973.

26. Hombert, J. M. Phonetic explanation of the development of tones from prevocalic consonants, in UCLA Working Papers in Phonetics, 33, 1976.

27. Hombert, J. M. Consonant types, vowel height and tone in Yoruba, in UCLA Working Papers in Phonetics, 33, 1976.

28. Hombert, J. M. Development of tones from vowel height, in UCLA Working Papers in Phonetics, 33, 1976.

29. Huggins, A. W. F. Just noticeable differences for segment duration in natural speech, Journal of the Acoustical Society of America,51.4, pp.1270-1278, 1972.

30. Huggins, A. W. F. On the perception of temporal phenomena in speech, Journal of the Acoustical Society of America,51.4, pp.1279-1290, 1972.

31. Hunnicutt, S. A new morph lexicon for English, manuscript from the Research Laboratory of Electronics, M. I. T., 1976.

32. Hyman, L. M. The role of consonant types in natural tonal assimilations, in Consonant types and tone (Ed. L. Hyman), Southern California Occasional Papers in Linguistics, No. 1, 1973.

33. Jackendoff, R. Introduction to the X-Bar convention. Bloomington, Ind: Indiana University Linguistics Club, 1974.

34. Joos, M. Readings in linguistics, 4th Edition, University of Chicago Press, Chicago: 1957:1968.

35. Klatt, D. H. Segmental duration in English, Journal of the Acoustical Society of America, 59.5, pp.1208-1221, May 1976.

36. Lakoff, G. The global nature of the Nuclear Stress Rule, Language 48.2, pp.285-303, 1972.

37. Langendoen, D. T. Finite-state parsing of phrase-structure languages and the status of readjustment rules in the grammar, Linguistic Inquiry 6.4, pp.533-554, 1976.

38. Leben, W. R. The tones in English intonation. Linguistic Analysis, 2.1, pp.69-107, 1976.

39. Lehiste, I. Rhythmic units and syntactic units in production and perception, Journal of the Acoustical Society of America, 54.5, pp.1228-1234, 1973.

40. Lehiste, I., J. P. Olive and L. A. Streeter, Role of duration in disambiguating syntactically ambiguous senterces, Journal of the Acoustical Society of America, 60.5, pp. 1199-1202, 1976.

41. Levitt, H. and L. R. Rabiner, Analysis of fundamental frequency contours in speech, Journal of the Acoustical Society of America, 49.2, pp. 569-582, 1971.

42. Liberman, M. Y. The intonational system of English, unpublished M.I.T. dissertation, 1975.

43. Liberman, M. Y. and A. Prince, On stress and linguistic rhythm, to appear in Linguistic Inquiry, 1977.

44. Liberman, M. Y. and I. Sag, Prosodic form and discourse function, Papers from the Tenth Regional Meeting Chicago Linguistic Society, Chicago: pp.416-28, 1974.

45. Levine, A. Report on prosodic generation, Institute

for Mathematical Studies in the Social Sciences, Technical report no. 272, Stanford University: June, 1976.

46. Makkai, V. B. Phonological Theory, Holt, Rineheart and Winston, Inc. New York: 1972.

47. Marcus, M. 'Wait and See' strategies for parsing natural language, Working papers 75, MIT-AI laboratory, 1974.

48. Marcus, M. A design for a parser for English, ACM'76, Proceedings of the Annual Conference pp. 62-68, October, 1976.

49. Martin, Ph. Analyse phonologique de la phrase francaise, Linguistics, 146.35-67, 1975.

50. Martin, Ph. Elements pour une theorie de l'intonation, Rapport D'activities de L'institut de Phonetique, Institut de Phonetique, Brussels, Belgium, R.A. 9/1, pp.98-126, Sept.'74-Mar.'75.

51. Melby, A. K., R. A. Millet, E. G. Lytle and W. J. Strong, Modifying fundamental frequency contours, paper presented at the 90th Meeting of the Acoustical Society of America, 1975.

52. Miller, G. A. The magical number seven plus-or-minus two; Some limits on our capacity for processing information, Psychological Review, 77, pp. 237-273. 1970.

53. Nakatani, L. H. Sentence prosody in isolated and in story contexts, The Journal of the Acoustical Society of America, 58 Supp. 1, 1975.

54. Newman, S. On the stress systems of English, Word, 2.171-87, 1946.

55. O'Connell, D. C., E. A. Turner, and C. A. Onuska, Intonation, grammatical structure and contextual association in immediate recall, Journal of Verbal Learning and Verbal Behavior 7, pp. 110-6, 1968.

56. O'Connor, J. D., and G. F. Arnold, Intonation of colloquial English, London: Longmans, 1961.

57. Ohala, J. Aspects of the control and production of speech, Working Papers in Phonetics 15, UCLA, April, 1970.

58. Ohala, J. The physiology of tone, in Consonant types and tone, ed. by L. Hyman, Southern California Occasional Papers in Linguistics, No. 1, 1973.

59. Ohman, S., and J. Lindqvist, Speech analysis and speech production, Quarterly Progress and Status Report, 4, Speech Transmission Laboratory, pp. 1-6, 1965.

60. Ohsiek, D. O. Accent and intonation in Japanese, paper presented at California Linguistic Society meeting, May, 1976.

61. Olive, J. P. Fundamental frequency rules for the synthesis of simple declarative sentences, Journal of the Acoustical Society of America, 57.2, pp. 476-482, 1975.

62. Olive, J. P., and L. H. Nalcatani, Rule synthesis of speech by word concatenation: A first step, Journal of the Acoustical Society of America, 55.3, pp. 660-666, 1974.

63. Oller, D. K. The effect of position utterance on speech segment duration in English, Journal of the Acoustical Society of America,54.5, pp.1235-1247, 1973.

64. O'Shaughnessy, D. Modeling fundamental frequency, and its relationship to syntax, semantics, and phonetics, doctoral dissertation, M.I.T. 1976.

65. Palmer, H. E. English intonation, with systematic exercises, Cambridge: Heffer, 1922.

66. Ross, J. R. Conditions on variables in syntax, doctoral dissertation, M.I.T. Reprinted by, Bloomington, Ind: Indiana University Linguistics Club, 1967.

67. Salzer, F. Structural hearing, New York: Dover Publications, 1962.

68. Sanders, W. R., G. V. Benbassat, and R. L. Smith, Speech synthesis for computer assisted instruction: The MISS system and its applications. SIGCUE Bulletin, 8.1, pp. 200-211, 1976.

69. Sanders, W. R., and G. V. Benbassat, The MISS speech synthesis system. Institute for Mathematical Studies in the Social Sciences, Stanford University, in preparation.

70. Sapir, E. Sound patterns in Language, Language, 1.37-51, 1925. Reprinted in Joos (1968) as pp. 19-25.

71. Sapir, E. The psychological reality of phonemes, 1933. Translated and reprinted in Makkai (1972) as pp. 22-31.

72. Schank, R., and K. Colby (eds.), Computer models of thought and language, Freeman Press, 1973.

73. Selkirk, E. O. The phrase phonology of English and French, unpublished M.I.T. doctoral dissertation, 1972.

74. Schenker, H. Harmony, O. Jonas (ed.), E. M. Borgese(trans.), University of Chicago Press, 1906:1954.

75. Schmerling, S. F. Aspects of English Sentence Stress, doctoral dissertation, University of Illinois, 1973. Reprinted by University of Texas Press, Austin, Texas.

76. Suppes, P., R. Smith and M. Beard, University-level computer-assisted instruction at Stanford: 1975, Technical report no. 265, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1975.

77. Swadesh, M. The phonemic principle, Language, 10.117-29, 1934. Reprinted in Joos (1968) as pp. 32-7.

78. Twaddell, W. F. On defining the phoneme, Language monograph No. 16, 1935. Reprinted in Joos (1968) as pp.32-7.

79. Wasow, T. Transformations and the lexicon, in P. Culicover, T. Wasow and A. Akmajian (eds.) Formal Syntax, Academic Press, 1977.

80. Williams, E. S. Underlying tone in Margi and Igbo, Linguistic Inquiry 7.3, pp.463-484, 1971:1976.

81. Wingfield, A. and J. F. Klein, Syntactic structure and acoustic pattern in speech perception, Perception and Psychophysics 9.23-25, 1971.